DOCTOR OF PHILOSOPHY

Rapid analysis of pharmacology for infectious diseases.

Carruthers, Ian Michael

Award date:
2013

DOCTOR OF PHILOSOPHY

Rapid analysis of pharmacology for infectious diseases.

Ian Michael Carruthers

2013

University of Dundee

Rapid analysis of pharmacology for infectious diseases.

Ian Michael Carruthers

Biological Chemistry and Drug Discovery

University of Dundee

A thesis submitted for the degree of

Doctor of Philosophy

March 2013

for Helen...

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported me during this project. First and foremost I would like to thank my supervisor, Professor Andrew Hopkins DPhil FRSC FSB, for the opportunity to study in his group, his countless support, encouragement and positivity that was given throughout. Special thanks go to Dr. Richard Bickerton, not only has he given me innumerous advice and guidance, but along with his wonderful wife Ruth, friendship and hospitality. My thanks go to Jeremy Besnard, Professor Paul Wyatt and Dr. Nick Leslie for useful discussions and feedback.

I would also like to extend my thanks to Pfizer and the EPSRC for their kind financial support.

Finally, I would like to thank all my family, especially Mum, Dad, Jacky, Jim and Lee. And to those who gave me the love and motivation to keep going; my darling wife Helen, beautiful son Sebastian and faithful dog Hendrix.

Declaration

The following work was carried out under the supervision of Prof. Andrew L. Hopkins at the College of Life Sciences, University of Dundee between April 2009 and March 2013. All references cited in the document have been consulted, unless otherwise stated. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This thesis has not been submitted in whole or in part for any degree, diploma or qualification at any other university.

The relevant Ordinance and Regulation have been fulfilled.

Abstract

Infectious diseases represent a multitude of threats to populations in both the developed and developing world, from the emergence of drug resistant bacteria and new pathogens to the ancient killers of the neglected topical diseases. Yet a common problem unites all infectious diseases, that is the challenge of how do we cost effectively identify new drugs? The arrival of high-throughput low cost sequencing starkly illustrates the nature of the challenge: the genome sequence of any pathogen can now be determined in a few days yet the availability of complete pathogen genomes has not led to the anticipated wave of new therapies. One reason for this failure might be that previous efforts at selecting the best targets from the genome have not taken into account information on the properties of associated small molecule ligands. To improve the exploitation of genomic information in the discovery of drug targets for new anti-infective agents a modular informatics framework is described that enables the large-scale comparative analysis of pathogen and host genomes. Specifically, new methods to predict essential genes, identify druggable domains and predict selectivity are presented, that have advantages over current approaches. The proposed method to predict essentiality is benchmarked against whole genome essentiality datasets and applied in practice to the anal-

ysis of a diverse range species including the bacterial pathogen *Pseudomonas aeruginosa* and eukaryotic parasites *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania braziliensis*, *Leishmania infantum*, *Leishmania major* and *Schistosoma mansoni*. In order to identify druggable and selective targets a domain-based approach to mining genomes for druggable targets is developed. A domain family based approach enables the determination of "binding site signatures" in the primary amino acid sequences which enables the identification and comparison of specific binding modes for both active/orthosteric site and allosteric site ligands. Information in the binding site signatures is used to train and validate a Bayesian model to predict a compounds selectivity between members of a domain family, whether from within a single genome or from multiple species.

Contents

List of Figures

Chapter 1

Introduction

"It is time to close the book on infectious diseases, and declare the war against pestilence won", is a quote often attributed to Dr. William H. Stewart, United States Surgeon General from 1965 to 1969 (Spellberg, 2008). Whilst this quote may be apocryphal, it has often been used to highlight the mistaken belief that pathogen-borne diseases are no longer a threat to human life. It may be surprising to some, that shortly after his discovery of penicillin in 1928, Alexander Fleming, having observed resistance develop in the laboratory gave a stark warning in his 1945 Nobel prize lecture: *"there is the danger that the ignorant man may easily underdose himself and by exposing his microbes to non-lethal quantities of the drug make them resistant"* (Fleming *et al.*, 1945). He was not wrong, indeed since the introduction of the penicillins in the late 1940's antibiotic resistance has rapidly followed, and as early as 1953 an outbreak of *Shigella dysenteriae* was found to show resistance to four different classes of antibiotics (Todar, 2008).

In the West, the incidence of multiple antibiotic resistant bacterial strains has steadily increased year on year. In the US, methicillin-resistant *Staphylococ-*

cus aureus, kills more people than emphysema, HIV, Parkinson's disease, and homicide combined (Idsa, 2011), and indiscriminately effects people of all ages. Bacterial drug resistance is becoming an increasing factor in hospital treatments (Figure 1.1), increasing length of stay, healthcare costs, morbidity and mortality (Davies & Davies, 2010).

Drug resistant strains are just some of the nearly 340 infectious diseases that have emerged across the globe since 1940 (Jones *et al.*, 2008). Globally, the World Health Organization (WHO) estimated that approximately one billion people suffer from at least one, and some more than one, neglected infectious diseases such as schistosomiasis, soil-transmitted helminthiasis, blinding trachoma, onchocerciasis, trypanosomiasis and lymphatic filariasis (Chan, 2007). The poorest populations of the world are also at risk from the major infectious diseases such as malaria, tuberculosis and human immunodeficiency virus (HIV). Half of the disease burden of 80% of the world's population, that reside in the developing countries is due to communicable diseases.

Despite the significant global problems associated with emerging and neglected infectious diseases, there remains a lack of progression within the pharmaceutical industry regarding new and novel therapies. In the last four decades only one novel class of antibiotic, the oxazolidinones, has been developed and licensed for treatment of gram positive bacterial infection (Herrmann *et al.*, 2008), and there remain few antibiotics in the development pipeline for gram negative species (Projan & Bradford, 2007). One reason for this limited success is shared with all drug discovery programmes - that the high attrition rate of compounds entering clinical development requires a huge financial investment. In the case of antibacterials, any successful drug revenue can be quickly diminished by the

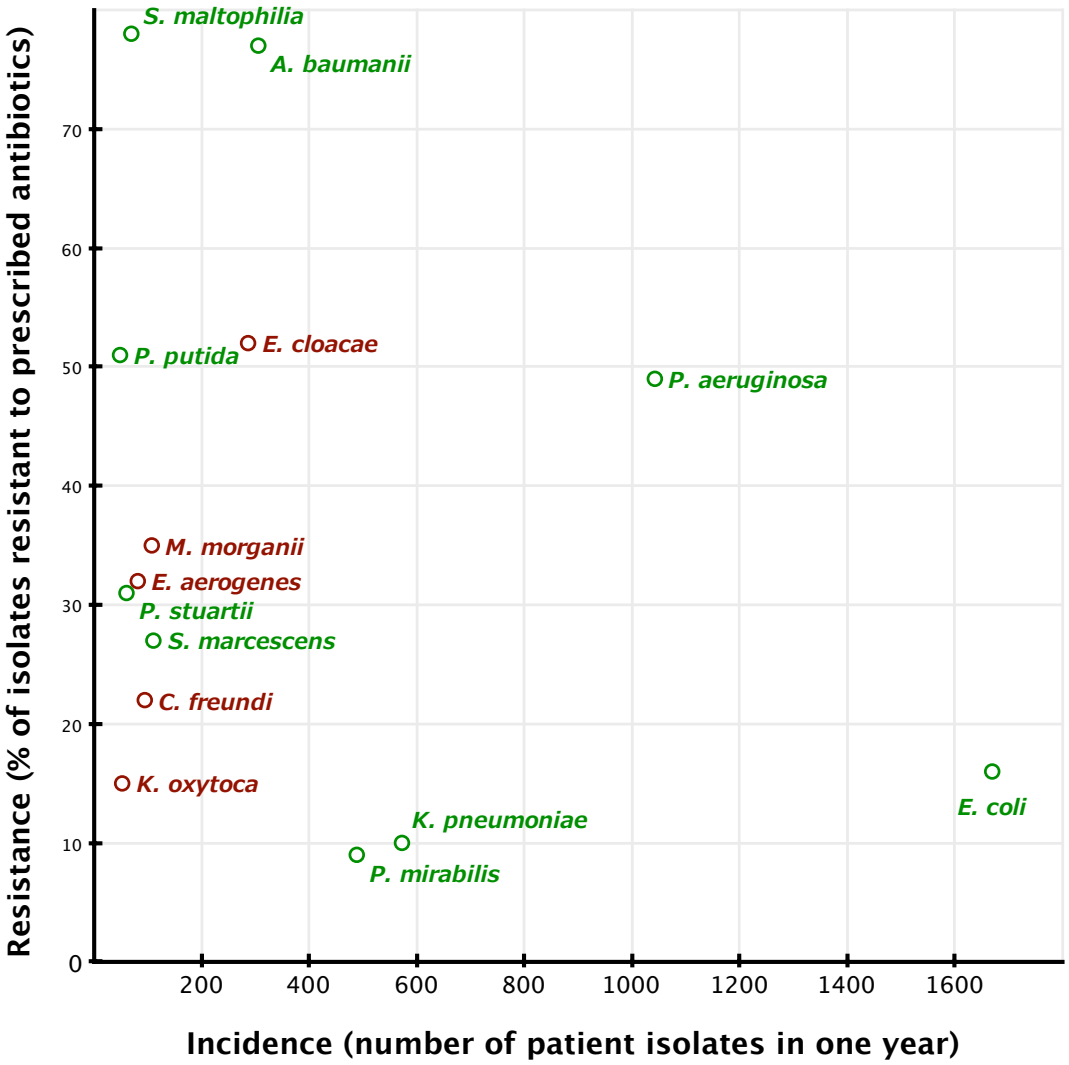


Figure 1.1: The incidence of infection, and levels of resistance observed to the routinely used antibiotics, in a major Memphis hospital (Richard Lee, 2009, St. Jude Children's Research Hospital. personal communication). The resistance shows the average ineffectiveness (%) of the currently prescribed antibiotics for the species. Where the species had a genome sequencing project ongoing or completed, in green. No genome project (in 2009) in red.

emergence of resistance, or held in reserve by physicians as a last resort, therefore limiting the revenue achievable during the patent term. In the case of neglected tropical diseases, the most affected populations are often the poorest, and as such investment costs are difficult to recoup. A common problem unites all infectious diseases whether they are pandemic, epidemic or endemic infectious diseases - that is the challenge of how do we cost effectively identify new drugs?

1.1 Post-genomic era

The impact of the genome sequences on infectious disease drug discovery has, to date, been disappointing. The outbreak of Shiga toxin-producing *Escherichia coli* strain O104:H4 in Europe, in the summer of 2011, demonstrated the genome sequence of an emerging pathogen can now be determined within a few days of its identification (Rohde *et al.*, 2011). The ability to swiftly examine a pathogen genome, has had little or no effect on the capacity to develop novel, target-led, anti-infective drugs (Payne *et al.*, 2006). The availability of complete pathogen genomes should enable systematic, rational prioritization of all potential drug targets for a given pathogen (White & Kell, 2004). However, the initial excitement that heralded the release of the first pathogen genomes has been tempered by the failure of the first generation genomics-led anti-bacterial drug discovery campaigns to yield the anticipated wave of new therapies (Livermore *et al.*, 2011; Payne *et al.*, 2006). One contributing factor for this disappointing outcome is that despite significant effort invested in understanding the biological considerations that underpin a good molecular target (Nagaraj & Singh, 2010; Payne *et al.*, 2004), there has been little effort to consider the equally important chemical

aspects, such as druggability, species selectivity, chemical diversity and the appropriateness of the chemical space accessed by compound libraries being screened (Brötz-Oesterhelt & Sass, 2010; Gwynn *et al.*, 2010; Livermore *et al.*, 2011).

1.2 Chemogenomics

The arrival of large-scale chemogenomic resources, such as ChEMBL (Gaulton *et al.*, 2011), provides an unprecedented opportunity to change this situation (Bellis *et al.*, 2011). Collectively, these resources provide for the first time something approaching a global view of pharmacological space (Paolini *et al.*, 2006). Harnessing this knowledge enables new systematic approaches to be developed to inform drug target selection (L. Hopkins *et al.*, 2011). Assessment of the chemical space associated with a particular target provides the means to make an indirect evaluation of the likelihood of binding drug-like chemical matter (Bickerton *et al.*, 2012) that is not dependent on the availability of a crystal structure. This chemogenomic druggability approach also carries the obvious advantage of suggesting potential chemotypes to seed future development. The question then becomes one of defining the most appropriate chemical space. As well as the obvious constraints on molecular structure imposed by the requirement for target-based activity, which will vary according to the target, other constraints on the physico-chemical properties will be imposed regardless of the target (Hopkins & Bickerton, 2010; O'Shea & Moser, 2008). These constraints may include the requirement for oral bioavailability or permeation across different cellular barriers (e.g. human gut, bacterial cellular envelope or the blood brain barrier for central nervous system penetration). Such constraints should be defined *a priori*, by

amongst other things, the Therapeutic Product Profile (TPP) (Curry & Brown, 2003; Frearson *et al.*, 2007; Wyatt *et al.*, 2011) and include considerations such as the route of administration, and the cellular and subcellular location of the molecular target or targets.

1.3 Target Selection

Target selection criteria include biological features such as gene essentiality (typically determined by large-scale knockout experiments) and selectivity (typically determined by an absence of equivalent protein in the human host). Such fundamental criteria remain important, but they must be augmented by an equivalent consideration of underlying chemical factors. In recent years there have been several attempts to prioritize drug targets from the genomes of pathogens: *Plasmodium falciparum* (Joubert *et al.*, 2009), *Schistosoma mansoni* (Berriman *et al.*, 2009; Caffrey *et al.*, 2009), *Mycobacterium tuberculosis* (Hasan *et al.*, 2006; Raman *et al.*, 2008; Singh *et al.*, 2006; White & Kell, 2004), *Vibrio cholerae* (Katara *et al.*, 2010), *Brugia malayi* (Kumar *et al.*, 2007), *Staphylococcus aureus* (White & Kell, 2004) and *Escherichia coli*(White & Kell, 2004). The most extensive attempt to prioritize neglected disease drug targets to date is the TDR Targets Database (Aguero *et al.*, 2008) that contains information on 12 pathogen genomes. However, even the TDR database has a somewhat limited approach to assessing target essentiality, which was logically intuitive, but had not been examined for accuracy.

Multiple criteria can be considered when selecting and prioritizing potential targets encoded in the genome: precedence, druggability, perturbation/essential-

ity, selectivity, and spectrum of activity. Given the finite number of potential targets (defined by the pathogen proteome), inclusion of additional criteria raises the quality and reduces the size of the resulting prioritized target pool. By taking a broad approach to defining each criterion and using the number and quality of evaluations to rank targets, the size and quality of the resulting pool can be maximized. Given the growing availability of complete pathogen genomes, the target selection and prioritization frameworks are applicable to any pathogen genome. Indeed, the same considerations are also pertinent to bacterial, protozoal, helminthic, viral or fungal pathogens of humans, animals or plants (L. Hopkins *et al.*, 2011).

1.3.1 Precedence

A natural starting point in any target selection strategy would be precedence - i.e. identification of the known drug targets. Aside from the obvious application of this set as potential targets themselves, their identification also aids analysis of the properties that differentiate them from non-drug-targets, to inform future target selection strategies.

1.3.2 Druggability

A range of definitions of druggability has been suggested (Keller *et al.*, 2006). The most widely used definition is: a druggable target is one that has the capacity to bind drug-like chemical matter. Importantly this definition is independent of the wider implications of modulation of the molecular target on cellular function and biology or issues around ligand selectivity. Most published approaches

consider druggability qualitatively - a target is classified as being druggable or otherwise (Halgren, 2009; Krasowski *et al.*, 2011). A more nuanced approach may be to consider druggability as a continuum, ranging from highly druggable molecular targets known to bind several different drug-like chemotypes to targets that have no characterized binding site or whose known ligands have unfavorable properties. Consideration of druggability in quantitative terms enables targets to be prioritized objectively and the desired number of high-ranking targets selected according to available capacity. Several approaches have been developed to assess target druggability, most of which use structural information to characterize ligand-binding sites. The open source fpocket algorithm (Le Guilloux *et al.*, 2009) employs Voronoi tessellation method to detect protein cavities which scored using a logistic model trained with three descriptors: local hydrophobic density, hydrophobicity and normalized polarity. Halgren (2009) describes the SiteMap package that uses a grid method incorporating van der Waals energies and a buriedness term to predict protein pockets which are assessed using a scoring function (Dscore) that includes terms for the pocket size, enclosure, as well as a penalty for its hydrophilicity. In DoGSiteScorer (Volkamer *et al.*, 2012) pockets and subpockets are predicted using a difference of Gaussian filter and druggability prediction made using a machine learning approach based on global descriptors and a nearest neighbor approach based on local features. In DrugPred, Krasowski *et al.* (2011) use a partial least-squares discriminant analysis that considers the size, polarity, and hydrophobicity of the binding pocket. Applying such structure-based approaches at genome scale carries the inherent limitation that an accurate protein structure is required. For example, in the case of the gram-negative bacteria *P. aeruginosa* this would limit their applicability

to 5.0% of the proteome. Whilst coverage could be extended through homology modelling approaches, the requirement for a highly accurate model of a ligand-binding cavity, including the correct orientation of amino acid side-chains, means that in all but the most straightforward of cases comparative models are likely to have limited use. Even when a structure is available the location of the most relevant binding site is not always known and is not necessarily obvious. While the presence of bound ligands can guide identification of the relevant site, many structures have bona fide binding sites that are unoccupied while others have multiple occupied binding sites. A further complication is that the ligands found in structures may include compounds used in the crystallization conditions such as buffers, detergents or other additives that are not of physiological or pharmacological interest. The visual inspection of potentially hundreds of structures is not practical in a high-throughput setting. A range of computational methods exist that can identify binding sites on protein structures but their accuracy (Bianchi *et al.*, 2012) limits their use in high-throughput automated studies. An alternative approach, independent of the requirement for a protein structure, would be to assess the quality of a target by analyzing the potency and drug-likeness of the chemical matter already associated with it. This chemogenomic druggability approach is expedited by the advent of large-scale databases of bioactivity that collectively associate millions of compounds with thousands of molecular targets. The European Bioinformatics Institute's ChEMBL database (Gaulton *et al.*, 2011) of bioactivity extracted from the medicinal chemistry literature is an important database resource for such analyses.

1.3.3 Perturbation

A clear requirement of a target is that its modulation leads to the inhibition, disruption or perturbation of the disease pathology at the cellular level. When looking for drug targets in a pathogen, conventionally this is addressed by considering genetically essential genes i.e. those that lead to a lethal phenotype when knocked out (Frearson *et al.*, 2007). Essentiality itself is inherently context dependent: genes that are dispensable in rich media may be essential in minimal media or *in vivo* (D'Elia *et al.*, 2009). The interpretation of such data should be restricted to the conditions under which the experiment was performed. In a broader sense it may be more appropriate to consider the fact that many validated drug targets are themselves not genetically essential (Denome *et al.*, 1999; Janoir *et al.*, 1996). Therefore, perturbant targets are those whose chemical modulation leads to a perturbation of the disease phenotype. Beyond genetically essential targets, perturbative targets may also include polypharmacology targets, chemically validated targets, virulence factors, host factors and synthetically lethal gene combinations.

Analysis of screening data suggests that only approximately 15% of a typical proteome exhibit any evidence of being potentially modulated by drug-like compounds (Hopkins & Groom, 2002). Similarly genome-scale knockout studies in model organisms have identified that generally <20% of genes are individually essential (Baba *et al.*, 2006; Kamath *et al.*, 2003; Winzeler *et al.*, 1999). Assuming for the purposes of illustration that druggability and essentiality are independent factors, and assuming 15% of targets are druggable and 20% of targets are essential; then the targets that are both lethal and druggable would only be the

intersect i.e. 2-3% of the proteome. Even for a bacterium with a relatively large proteome like *P. aeruginosa* this leaves only around 100-150 potential targets. Further, when other important criteria are also considered, such as selectivity and activity spectrum, then even this small percentage of suitable proteins is further depleted. The result of these issues is that the set of targets that can be identified, fulfilling all of the relevant criteria (i.e. the "opportunity space") may number just a few dozen distinct proteins. In order to maximize this set, the approach should be to use multiple orthogonal methods and to integrate the results to increase coverage. Confidence in each individual target is enhanced when multiple independent methods give corroborating predictions, thereby providing the means to rank targets.

1.3.3.1 Polypharmacology

Polypharmacology describes the property of some drugs to interact with multiple proteins simultaneously. Often this feature is unwanted, and where this occurs in drugs, it can cause adverse patient side-effects. Polypharmacology can also be beneficial, especially in anti-infectives, where the limited target space can be increased by disrupting multiple protein functions simultaneously. Where the targets are individually non-essential, due to pathway or functional redundancy, the combinations of functional inhibition may be fatal (Hopkins *et al.*, 2011).

This polypharmacology strategy can have the inherent advantage of reducing the rate of resistance developing via mutation, as the probability of mutations occurring in multiple genes simultaneously, is orders of magnitude higher than that of a single gene. Considering the disadvantage of the single target approach, it is perhaps no surprise to find that a large number of currently available an-

tibiotics operate on multiple proteins simultaneously. For example, the β-lactam antibiotics inhibit multiple related penicillin-binding proteins (PBPs) expressed by bacterial species. Due to overlapping and functional redundancy, none of these PBPs are individually essential (Denome *et al.*, 1999), but when multiple PBPs functions are inhibited the consequences are lethal.

The targeting of two or more essential genes, with a single drug, may be one strategy to delay the emergence of drug resistance. Alternatively, a drug that targets essential genes across a number of pathogens may have the potential to be a broad-spectrum anti-infective, assuming it is selective over any human homologs. Moreover, the search for broad-spectrum anti-infectives may be one strategy to help to reduce the burden of poly-parasitism (i.e. patients suffering from multiple parasitic infections), a condition which is common in the developing world (Pullan & Brooker, 2008). Given the difficulties associated with developing a potent drug against just one target, the chances of discovering a broad-spectrum agent for either bacterial or parasitic diseases may be very small. However, the opportunity provided by large-scale comparative chemogenomics across multiple genomes provides us with a means to search the landscape of infectious disease drug targets, for such drug targets may be essential and druggable yet common between several species.

1.3.4 Selectivity

Selectivity is a measure of a drugs ability to bind some proteins preferentially to others. In the case of anti-infectives, the requirement to bind and inhibit pathogen targets, but not evolutionary related host proteins. While selectivity

is often a metric of the drug, pairs of related proteins can be predisposed to bind similar chemical matter, due to similar properties of their ligand binding pockets. To combat potential pathogen-host selectivity issues, a conventional strategy is to consider only those targets that share no evolutionary equivalent with the host. Whilst such a requirement ensures selectivity, it also rules out many conserved gene products that may otherwise make tractable and valued targets, e.g. dihydrofolate reductase or the ribosome, which are both successful anti-infective targets (Silver, 2011) despite having human orthologs. Selectivity may be achieved by several means, but from a genomic perspective, the identification and exploitation of amino acid differences in the human-host binding sites offers the most feasible route.

1.3.5 Spectrum of activity

The presence of orthologous proteins in related pathogenic species opens up the possibility that successful compounds may exhibit broad-spectrum activity. The obvious clinical benefits of such an outcome would be tempered by the drawback that this could increase the selection pressure on gut flora and may therefore increase the likelihood of resistance occurring, and being spread via horizontal gene transfer. As such, targets having homologs or orthologs in related Gram-negative pathogens can be included or excluded depending on the desired strategy.

1.4 Aims

Highly efficient new approaches to infectious disease drug discovery are urgently required to face the global health challenges posed by emerging drug-resistant and

neglected infectious diseases. The ready availability of the pathogen and parasite genomes and the massive reduction in speed and cost of high-throughput genome sequencing now means that one can approach nearly every infectious disease with a knowledge of its genome and predicted proteome. Therefore the aim is to consider informatics-based methods that can help rapidly analyze genomes for potential drug targets by systematic comparative genomics, chemogenomics and structural bioinformatics.

1.4.1 Modular Implementation of Analysis

To identify potential drug targets from any genome, a range of informatics services are required to enable the range of diverse drug discovery approaches. The overall goal is to create informatics services or modules that can systematically infer the attributes of pathogen proteins. A basis for designing such an informatics strategy is to consider the common hypothesis of anti-infectious drug discovery which propose that a good anti-infective drug target need to satisfy several criteria:

- **Essential** - the target occupies a point in the cellular network whereby its modulation will disrupt function at the cellular level, with lethal consequences for the pathogen
- **Selective** - the drug should preferentially perturb the pathogen target over any human target.
- **Druggable** - the target has the capacity to bind and to be modulated by a drug-like small molecule.

In this thesis I propose and implement an infectious disease informatics strategy by a modular approach to drug target identification and prioritization. Each module addresses one of the hypothesized criteria. The intent is that each module has the ability to be applied generically to any of the genomes of interest.

The successful implementation of the modules would be achieved by fulfilling the following criteria:

- Identification of Essential genes in pathogens
  - Identify essential pathogen genes data.
  - Utilize genome-to-genome orthology mapping tools for transfer of essentiality information between genomes.
  - Utilize and, if possible, improve existing methodology for *ab initio*, *in silico* genome essentiality prediction.
  - Design a rigorous testing procedure to determine the predictive power of any essentiality prediction method.
- Prediction of Druggable Genes in pathogens:
  - Use precedence-based methods to prioritize targets homologous to proteins with known drug-like inhibitors.
  - Refine existing resources such as ChEMBL to associate data with specific ligand-binding sites rather than to whole proteins.
- Analysis of Binding site selectivity

- Understand the relationships between ligand-binding site properties and their relationship to the selectivity/potency of small molecule inhibitors.
- Catalogue observed binding sites of a protein family, for a binding-site ontology.
- Prediction of potential selectivity issues between human proteins and human pathogen targets being assessed for druggability.
- Highlight similar binding sites within a pathogen genome for potential polypharmacology targets, reducing the risk of resistance and increasing the "essential" space.

Chapter 2

Phylogenomic inference of essentiality

2.1 Introduction

2.1.1 Is essentiality required for drug targets?

The requirement that the modulation of an anti-infective drug target should lead to the perturbation of the disease pathology, has caused the focus of research on genetically essential targets. While genetically essential targets are important, other sources of perturbative targets are available.

Drugs that reduce the pathogen fitness can give the host immune system additional time to respond to infection. One such mechanism is the bacteriostatic drugs, that target proteins that are not always essential for cell-life, but required for the normal replication rate. The class of antibiotics known as the sulfonamides, prevent the action of the bacterial dihydropteroate synthase (Lopez *et al.*, 1987). This enzyme is required for the synthesis of folate, which in turn is required for DNA replication. This inhibition of DNA replication halts the bacterial cell cycle and stops the multiplication of bacteria, but does not kill them. The antibiotic trimethoprim targets bacterial dihydrofolate reductase (Hitchings, 1973, 1989), another enzyme involved in folate metabolism and therefore bacterial multiplication. Interestingly, both of the genes encoding these two enzymes (*folA* and *folP*) were classified as essential in *Pseudomonas aeruginosa* (Table 2.1), despite being known to be only bacteriostatic. This could be explained by the experimental methodology, which required gene-knockout mutants to form bacterial colonies to be classified non-essential. Therefore it is important to note that targeting genes classed as essential, may not kill the pathogen, but only prevent cell-division, which may not prevent the disease.

Polypharmocology is a strategy to target multiple proteins simultaneously,

therefore reducing mutation-bourn resistance and increase the essential target space. The cephalosporins are β-lactam antibiotics which are bactericidal by inhibiting penicillin-binding proteins (PBPs). Table 2.1 shows the five PBPs of *P. aeruginosa* (*ponA*, *mrcB*, *pbpA*, *ftsI* and *pbpC*), which are all individually non-essential. However, many cephalosporins including cefepime, are known to be effective against *P. aeruginosa* (Denome *et al.*, 1999) by simultaneously inhibiting multiple PBPS (Chapman & Perry, 2012).

The fidelity of viral and bacterial DNA replication is low, and bacterial and viruses can rapidly develop mutations that reduce drugs binding their proteins. An alternative approach is the targeting host cofactors that are often indispensable for the colonization and propagation of pathogens (Khattab, 2009; Vaudaux *et al.*, 1989). There has been some success with this approach for viruses, and recently the FDA (U.S. Food and Drug Administration) approved Maraviroc, an anti-retroviral drug that targets a human protein (C-C chemokine receptor type 5), to block HIV (Human immunodeficiency virus) infection of host cells (Friedrich *et al.*, 2011). This approach has also been advocated for pathogenic bacteria and protozoa (Prudencio & M. Mota, 2012; Schwegmann & Brombacher, 2008).

Virulence factors are genes not usually essential in bacteria, but are often associated with infectious disease causing strains. Many species such as *Clostridium botulinum*, *Escherichia coli* and *Staphylococcus aureus* do not usually cause disease, but the acquisition of genes from virulence factor-encoding bacteriophages can transform them into highly virulent pathogens (Keen, 2012). The targeting of these virulence factors with therapeutics offers the chance to nullify the pathogenicity of these species. Whilst there exists many classes of perturbative

targets, the genetically essentials are still an important and trusted source of anti-infective targets.

PA ID	Gene	Gene product	Essentiality
PA5045	*ponA*	penicillin-binding protein 1A	non-essential
PA4700	*mrcB*	penicillin-binding protein 1B	non-essential
PA0378		probable transglycosylase	non-essential
PA4003	*pbpA*	penicillin-binding protein 2	non-essential
PA4418	*ftsI*	penicillin-binding protein 3	non-essential
PA2272	*pbpC*	penicillin-binding protein 3A	non-essential
PA3047		probable D-alanyl-D-alanine carboxypeptidase	non-essential
PA3999	*dacC*	D-ala-D-ala-carboxypeptidase	non-essential
PA0869	*pbpG*	D-alanyl-D-alanine-endopeptidase	non-essential
PA4110	*ampC*	beta-lactamase precursor	non-essential
PA5514		probable beta-lactamase	non-essential
PA5302	*dadX*	catabolic alanine racemase	non-essential
PA4930	*alr*	biosynthetic alanine racemase	non-essential
PA4201	*ddlA*	D-alanine-D-alanine ligase A	non-essential
PA4410	*ddlB*	D-alanine–D-alanine ligase	non-essential
PA3168	*gyrA*	DNA gyrase subunit A	essential
PA0004	*gyrB*	DNA gyrase subunit B	non-essential
PA4964	*parC*	topoisomerase IV subunit A	essential
PA4967	*parE*	topoisomerase IV subunit B	essential
PA4450	*murA*	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	potential essential
PA4238	*rpoA*	DNA-directed RNA polymerase alpha chain	essential
PA4270	*rpoB*	DNA-directed RNA polymerase beta chain	essential
PA4269	*rpoC*	DNA-directed RNA polymerase beta′ chain	non-essential
PA4462	*rpoN*	RNA polymerase sigma-54 factor	non-essential
PA5337	*rpoZ*	RNA polymerase omega subunit	non-essential
PA0350	*folA*	dihydrofolate reductase	essential
PA4750	*folP*	dihydropteroate synthase	essential
PA4560	*ileS*	isoleucyl-tRNA synthetase	essential
PA1806	*fabI*	NADH-dependent enoyl-ACP reductase	non-essential
PA0286	*desA*	delta-9 fatty acid desaturase, DesA	non-essential
PA4266	*fusA1*	elongation factor G	non-essential
PA2071	*fusA2*	elongation factor G	non-essential

Table 2.1: Known antibacterial drug target equivalents in *Pseudomonas aeruginosa* (Silver, 2011). The individual essentiality of each gene assessed by Jacobs *et al.* (2003). Note that not all of these will be the targets of drugs used clinically against *P. aeruginosa* itself.

2.1.2 Experimental methods to identify essential genes

Experimental techniques to identify essential genes *in vitro* are well developed and documented (Gaiano *et al.*, 1996; Glass *et al.*, 2006; Kempheus, 2005). Transposon mutagenesis involves inserting a genetic element into the coding- or promoter-region of a target gene, thus potentially disrupting its function (Akerley *et al.*, 1998). Another method is to insert an inducible promoter region upstream of the target gene so gene expression may be activated on demand (Guzman *et al.*, 1995). For more complex organisms such as eukaryotes, anti-sense RNA may be inserted into the cells to bind targeted messenger RNAs (mRNA) thus reducing translation of gene products (Weiss *et al.*, 1999). RNA interference (RNAi) is a method, in which double-stranded RNAs (dsRNA) are introduced into the cell, and trigger the host RNAi pathway to selectively degrade mRNAs (Harborth *et al.*, 2001). The RNAi pathway is present in many eukaryotic species, but some species such as *Leishmania major* lack the RNAi mechanism (Robinson & Beverley, 2003). Each of these methods have their own advantages and problems. Two problems shared by all the experimental techniques, are the cost and time. These methods require extensive laboratory work, which may be appropriate for a select few high value targets, or a small number of whole organisms, but has been historically unfeasible for routinely and rapidly, defining the essential gene compliment of whole genomes. Rapid genome-wide approaches have recently been developed such as the RIT-seq system (Alsford *et al.*, 2011), used to screen *Trypanosoma brucei*, and the TraDIS system (Langridge *et al.*, 2009), used to screen *Salmonella enterica* serovar Typhi. Despite these rapid screening methods, the genomes of many hundreds of species have been fully sequenced,

but only a small percentage have genome wide essentiality data experimentally derived. It is the disconnect, between the speed of genome sequencing, and the speed of experimental determination of essential genes, that drives the work in this chapter.

2.1.3 *In silico* methods to identify essential genes

2.1.3.1 Homology based prediction methods

Similar protein sequences, often have similar functions (Duan *et al.*, 2006), and a standard informatics technique to transfer annotation from one sequence to another is by considering homology. Two sequences are homologous to each other if they share a common ancestor. This homology can be predicted by considering the detectable and statistically significant similarity of two sequences. Novel sequences can be searched against databases of known essential genes and where related sequences are observed, essentiality can be inferred. A recent study by Holman *et al.* (2009) showed that a variation of this method performed well at ranking known essential genes over known non-essentials in many species. It was noted that certain species (*Haemophilus influenzae*, *Helicobacter pylori* and *Escherichia coli*) were less amenable to this method of essentiality prediction than others. This was attributed to multiple factors, the performance of such methods will be influenced by many species specific factors, the proportion of essential genes within the species, the number of genes from closely related species reported in the database; and many other factors related to the biological niche of the species.

2.1.3.2 Orthology based prediction methods

Mushegian & Koonin (1996) describe a method to identify the minimal essential bacterial genome, based on orthology to a distant relative species. The reasoning was that those genes which were conserved after speciation, despite a large evolutionary time and pressure on genome size, are likely to be essential. At the time, the computational methods available for orthology detection were poor, and there was little experimentally derived essentiality data available. More recently Aguero *et al.* (2008) and many others (Caffrey *et al.*, 2009; Kumar *et al.*, 2007), have used orthology to infer functional characteristics such as essentiality across genomes. This method is attractive, as orthology detection capabilities have improved substantially, as well as the rate of genome sequencing and essential data availability. Intuitively it may be expected that, an essential gene in one species is likely to be essential in a closely related species, however, historically the application of this reasoning has presented some problems. Using a defined cutoff, as to what passes the criteria for essential, will invariably lead to more false negatives if the criteria are too strict (a lack of sensitivity), and false positives (a lack of specificity) if the criteria are too loose. Without a rigorous benchmark, it is impossible to understand how variations in prediction methods and criteria effect the accuracy of the results. Doyle *et al.* (2010) performed a benchmark of their own methods in predicting essential genes in eukaryotes, but there is a lack of such evidence for prokaryotes. Here it is intended, that these homology and orthology based prediction methods are extended, as well as benchmarked to understand the level of confidence in the predictions.

2.1.4 Understanding homologs, orthologs and paralogs

2.1.4.1 Homologs

Nature is a tinkerer and not an inventor (Jacob, 1977). New genes, and by extension proteins, evolve by duplication of existing genes, and mutations within the DNA sequence give rise to differences within the amino acid sequence and thus the properties of the resulting folded protein. These duplication and mutation events can, and do, occur many times, resulting in large families of genes which share a common ancestral gene. In general, the more recent the duplication event, the greater the similarity of the resultant amino acid sequences. However evolutionary distant, duplication events can still retain enough sequence similarity to be detectable by techniques such as the BLAST (Basic Local Alignment Search Tool)(see 2.2.5.1) (Altschul *et al.*, 1990). We can say that any pair of proteins that have arisen with characteristics inherited from a common ancestor, irrespective of the number of duplication events are "homologous" to one another.

2.1.4.2 Orthologs

When a species diverges into two distinct species (speciation), the equivalent genes within the two species are called "orthologous" to one another (Fitch, 1970). By definition, orthologs are also homologs. While these genes may undergo amino acid sequence divergence after speciation, in general, orthologous genes share the same or a similar functional purpose within the two species (Mushegian & Koonin, 1996).

2.1.4.3 Paralogs

When a gene duplication event occurs, the resulting two genes are said to be "paralogous" to each other. The new gene is also a paralog to any orthologs of the parent gene. If the duplication occurred after a speciation, the new gene is a called "in-paralog" of the existing gene. If the gene duplication occurred before speciation event, the resulting genes are called "out-paralogs". Paralogous genes have less evolutionary pressure to retain the same function as the parent gene and often diverge more rapidly than orthologs.

2.1.4.4 Co-Orthologs

When one or more duplication events occurs to an in-paralog, the resulting genes are collectively referred to as co-orthologs of the original ortholog in the other species. As these events can occur multiple times in either or both species, the resulting gene relationships can be one-to-one, one-to-many, or many-to-many.

2.2 Materials and methods

2.2.1 Database tables and loading

Relational databases (Codd, 1970) consist of a collection of interconnected sets of data stored in indexed tables. Interaction with the database is mediated by a relational database management systems (RDBMS), with SQL (Structured Query Language) - a standardized language for the addition, modification and retrieval of data, the creation and alteration of schema (the tables, constraints and relationships), and the management of database access. Oracle (http:

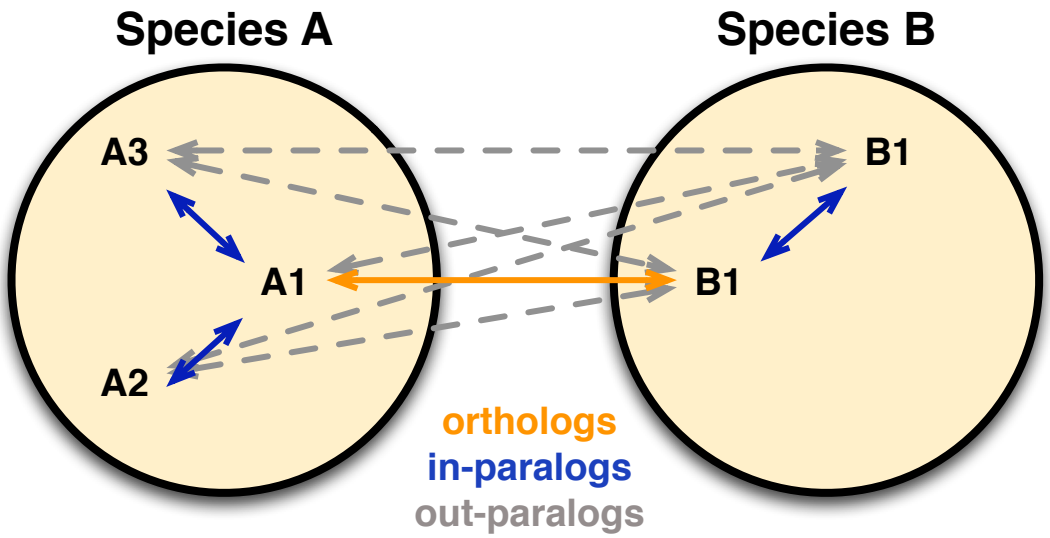


Figure 2.1: Ortholog relationships in two related species. The ancestral species had only gene A1, after speciation only genes A1 and B1 existed. All other genes occurred after the speciation. Orange arrows link orthologs. Blue arrows link in-paralogs. Broken gray arrows link out-paralogs. All arrows indicate co-ortholog pairs. All genes are homologs of all other genes. Adapted from Chen *et al.* (2007)

//www.oracle.com) is an industry standard RDBMS, and has been routinely demonstrated to provide good performance and reliability. There is excellent support software and API's (application programming interface) for interacting with Oracle. The benefits of relational databases are multiple. A well designed schema will enable complex data relationships to be modelled and enforced, make *ad hoc* queries fast, the ability to deal effectively with data growth (both in volume and type), and enable centralized access to data for multiple users. The essentiality components of the larger project database are shown in Figure 2.2. Proteome data was obtained from ftp.ncbi.nlm.nih.gov/genomes/Bacteria/. Taxonomy data was obtained from http://www.uniprot.org/taxonomy/ (Phan *et al.*, 2003). Experimental essentiality data is described in 2.2.2 and 2.2.3, and orthology data calculated as described in 2.2.6.1.

Throughout the work described below, data was inserted, manipulated and queried into the database using widely available tools such as Perl::DBI (http://dbi.perl.org/), cx_Oracle (cx-oracle.sourceforge.net and SQL*loader (http://www.orafaq.com/wiki/SQL*Loader_FAQ).

2.2.2 Gene essentiality data

A number of gene essentiality experiments have been performed and published, the Database of Essential Genes (DEG)(Zhang *et al.*, 2004) attempts to collate the results of these studies. The structure of DEG is such that it only records known essential genes from a species, and omits the known non-essentials. This is important as genes which are omitted from DEG maybe non-essential or may not have been assessed for essentiality. The Online GEne Essentiality (OGEE)(Chen *et al.*,

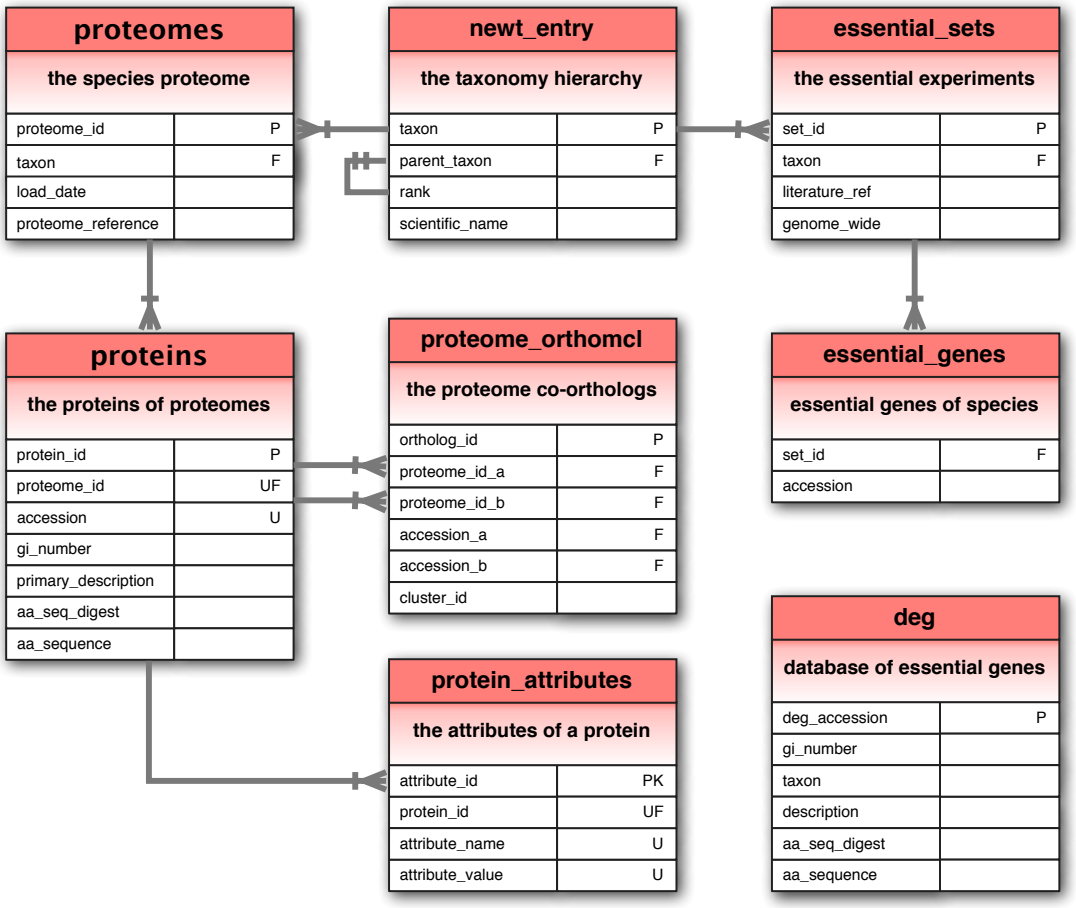


Figure 2.2: Database schema for proteome, essentiality and orthology data. The primary key, foreign keys and unique keys are represented by P,F and U respectively. Arrows indicate the direction of foreign key inheritance. For more details on column names and example queries see Appendix A.1. Figure generated using OmniGraffle (Case, 2013).

2012) database attempts to address this issue by recording non-essentials, but was unavailable at the time of this study. The DEG database was obtained from http://tubic.tju.edu.cn/deg/deg.rar, and contained 5260 essential genes, from 14 bacterial species (see Table 2.2).

Scientific name	NEWT taxon	Essential genes	Proteome size
Acinetobacter sp. ADP1	62977	499	3307
Bacillus subtilis str. 168	224308	271	4105
Escherichia coli K-12 MG1655	511145	712	4149
Francisella novicida U112	401614	392	1719
Haemophilus influenzae Rd KW20	71421	642	1657
Helicobacter pylori 26695	85962	323	1576
Mycobacterium tuberculosis H37Rv	83332	614	3989
Mycoplasma genitalium G37	243273	381	475
Mycoplasma pulmonis UAB CTIP	272635	310	782
Pseudomonas aeruginosa UCBPP-PA14	208963	335	5892
Salmonella typhimurium LT2	99287	230	4525
Staphylococcus aureus N315	158879	302	2619
Streptococcus pneumoniae	1313	244	1914
Vibrio cholerae	666	5	3,870

Table 2.2: Database of essential genes (DEG) summary. Only essential genes are available in DEG, those genes tested to be non-essential are not reported. Multiple copies of the same gene from each species may be present in DEG, if gene essentiality corroborated in multiple publications. (The specific revision of the genome used for the essential experiments and the protein prediction may differ.)

2.2.3 Genome-wide essentiality data

In order to benchmark essentiality prediction methods, it is necessary to have whole genomes with experimentally derived essentiality data available for the majority of the genes. Whole genomes are important, as most orthology detection software rely upon this. Genome-wide essentiality data ensures we know not only the essential genes, but the non-essential also, which enables accurate analysis of the predictions. For the purposes of the benchmark, only genomes with >90%

gene knockout (or equivalent) coverage were utilized. Five sets of published, genome wide, experimentally verified, essential genes were identified (Table 2.3). These genomes and associated annotations were extracted from the literature and added to the database (Figure 2.2).

The five essential genomes were experimentally measured in differing ways, and therefore the definition of an essential gene depended on the conditions and methodology used to determine them. In the methodology for these five bacterial species, all mutants were required to undergo several cycles of division before the non-essential genes could be detected. As a result of this division cycle, the observed essential genes may not have been essential for bacterial life, but just for reproduction. In such cases, targeting these genes with a therapeutic may have a bacteriostatic effect rather than a bactericidal effect.

Four species (*E. coli*, *F. novicida*, *M. genitalium* and *M. pulmonis*) were tested for essential genes required for viability on a rich medium, and one (*Acineobacter sp.* ADP1) on a minimal medium. The observed essential sets were smaller on rich media (≈300-400 genes) than on minimal media (499 genes), as would be expected, as those genes involved in the biosynthesis of essential compounds such as amino acids would be absolutely required on a minimal medium (de Berardinis *et al.*, 2008).

Three of the essentiality experiments (*F. novicida*, *M. genitalium* and *M. pulmonis*) were performed using random transposon mutagenesis, which relied upon a significant transposon saturation level. Where no transposon-mutant was recovered for a gene, then that gene was considered essential. A drawback to this method was that genes (especially small genes) may not be transposed by chance, and may have incorrectly been assigned as essential (Gallagher *et al.*, 2007). A

second drawback to this method was that those genes that were disrupted causing reduced fitness (not essential for survival), could have been out-competed by other mutants in the growth stages, and incorrectly deemed essential (Gerdes *et al.*, 2006).

The other two essentiality experiments (*E. coli* and *Acineobacter sp.* ADP1) were performed using gene-by-gene -deletion and -transposon mutagenesis respectively. The advantage of this method was that genes were not un-disrupted and predicted as essential by chance alone, and that mutant strains were cultured separately reducing competition.

Therefore, the definition of "essential genes" in the five genome-scale studies varied from genes "required for survival", genes "required for survival in a favorable environment" and genes "required for competitive growth" (Gerdes *et al.*, 2006). For a pathogenic bacteria, the infection process usually occurs in a nutrient rich environment (Rohmer *et al.*, 2011), and so this must be considered when using essential genes predicted on minimal medium. The gold standard for essentiality prediction in bacteria is a gene-by-gene deletion approach, as random transposon mutagenesis is often over-predictive of essential genes (Gallagher *et al.*, 2007). Despite the advantages of certain methodologies, those essentiality sets produced by potentially inferior methods were not ignored, as each bacterial species may have essential genes absent from other essential sets that could offer unique insight on a pathogens essential genes.

Scientific Name	Source	PSize	Tested	%	Essential	%	Growth media	Experimental
Acinetobacter sp. ADP1	de Berardinis *et al.* (2008)	3307	3195	96.6	499	15.6	min.	gene-by-gene transposon mutagenesis
Escherichia coli K12	Baba *et al.* (2006)	4390(4149)[1]	4288	97.7	300	7.0	rich	single gene-by-gene deletion
Francisella novicida U112	Gallagher *et al.* (2007)	1719	1720[2]	100	391	22.7	rich	random transposon mutagenesis
Mycoplasma genitalium G37	Glass *et al.* (2006)	475	475	100	381	79.6	rich	random transposon mutagenesis
Mycoplasma pulmonis UAB CTIP	French *et al.* (2008)	782	782	100	310	39.6	rich	random transposon mutagenesis

[1]In brackets represents the revised proteome size at the time of this work.
[2]discrepancy due to re-classified pseudogene.

Table 2.3: Whole genome essentiality screens data. PSize refers to the number of protein found in the species (at the time of the publication). Growth media describes the media on which knockouts/mutants were viable, min. refers to the minimal medium required for growth. Rich medium refers to a medium that would support a wide variety of bacterial species, and includes an amino acid source.

2.2.4 Existing orthology detection methods

When presented with two genomes or proteomes of distinct species, there is no guaranteed method to classify the orthologous and paralogous relationships. One of the reasons for this is that the two species may be separated by multiple speciation events, and the intermediate species extinct or unknown. The most reliable way of defining a pair of orthologs is to exhaustively assign biological functions to every protein in each genome, two genes which have the same function and share a solid underlying sequence relationship (homology) are most likely orthologs. Even with this process it is not always easy to distinguish between orthologs and out-paralogs. Using this manual method is of course impractical for even one pair of genomes, and impossible for rapidly detecting relationships between multiple genomes. Fortunately many algorithms and software have been developed for automatically detecting the orthologous relationships within two genomes. Many of these methods were exhaustively benchmarked for accuracy by Chen *et al.* (2007). The methods that attempt to recognize the problem of co-orthology generally perform better. The two best methods determined were INPARANOID (Berglund *et al.*, 2008) and OrthoMCL (Li *et al.*, 2003). While these two methods are comparable in overall accuracy, they differ in their prediction sensitivity and specificity, with OrthoMCL having a slightly lower false-negative rate than INPARANOID (0.07 vs 0.17), at a small cost to the false-positive rate (0.16 vs 0.07). OrthoMCL was deemed the better option, as by starting with more ortholog relationships, even at the expense of more incorrect relationships, would give a greater chance of predicting larger numbers of essential genes.

2.2.5 Benchmarking homology for essentiality inference

Each of the five benchmark genomes was compared individually versus the remaining four genomes, and against the DEG database using BLAST (section 2.2.5.1) to find essential homologs. The hypothesis being, that any protein that shared a common ancestor with a known essential protein, was probably essential itself. A simple variable metric, the percent coverage of the known essential protein sequence by the BLAST alignment, was used to distinguish significant hits. The expectation was that the greater the coverage of the essential protein, the more probable that any protein domain(s) required for the essential function, were maintained in the query protein, and therefore the essential function was also maintained.

2.2.5.1 Homology searches (BLAST)

The standalone version of the NCBI BLAST+ program (Camacho *et al.*, 2009) was obtained from `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.21/`, and installed locally. All databases builds and searches were performed with default parameters and an E-value cutoff of 1×10^{-03}, unless otherwise stated, as described at `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.21/user_manual.pdf`.

2.2.6 Benchmarking orthology for essentiality inference

Each of the five benchmark genomes was compared individually verses the remaining four genomes using OrthoMCL to establish co-orthology relationships. The distribution of co-orthology topologies can be seen in Figure 2.3. Direct 1:1

orthology relationships dominate, coving 66% of the relationships, a 1:0 topology (i.e. the gene has no detectable orthologs) represents 25% of the relationships, and with 1:2 (a single gene duplication after speciation in one species) covering 5%. The remaining 4% of the data being represented by numerous alternative topologies at low frequency.

2.2.6.1 OrthoMCL and parameters

OrthoMCL version 1.4 was downloaded from http://orthomcl.org/common/downloads/software/unsupported/v1.4/ , and installed locally. The program was modified to use the standard BLAST installation (see 2.2.5.1). The default parameters for e-value and MCL inflation index were used, as tightening these parameters only increases specificity at a larger cost on sensitivity (Chen *et al.*, 2007), while loosening the parameters reduces the granularity of the co-ortholog clusters, resulting in increased many-to-many topologies (Li *et al.*, 2003). All protein sequences representing the genomes, were taken from the Proteome database (Figure 2.2). OrthoMCL predicts the phylogenic relationships between proteins from a pair of species, however no phylogenetic tree can be constructed and so the true orthology relationships are only inferred. The outputs of OrthoMCL are clusters of proteins within the two given genomes representing co-orthologous groups, where the clustering produces a single gene from each genome, these genes are both considered orthologs. Where a cluster contains one gene from a species but multiple genes from the other species, the single gene is considered the ortholog, and the multiple genes considered paralogs. Where a cluster contains multiple genes from both genomes, all of these genes are considered paralogs to each other. Resulting orthology relationships were added to the proteome_orthomcl table of

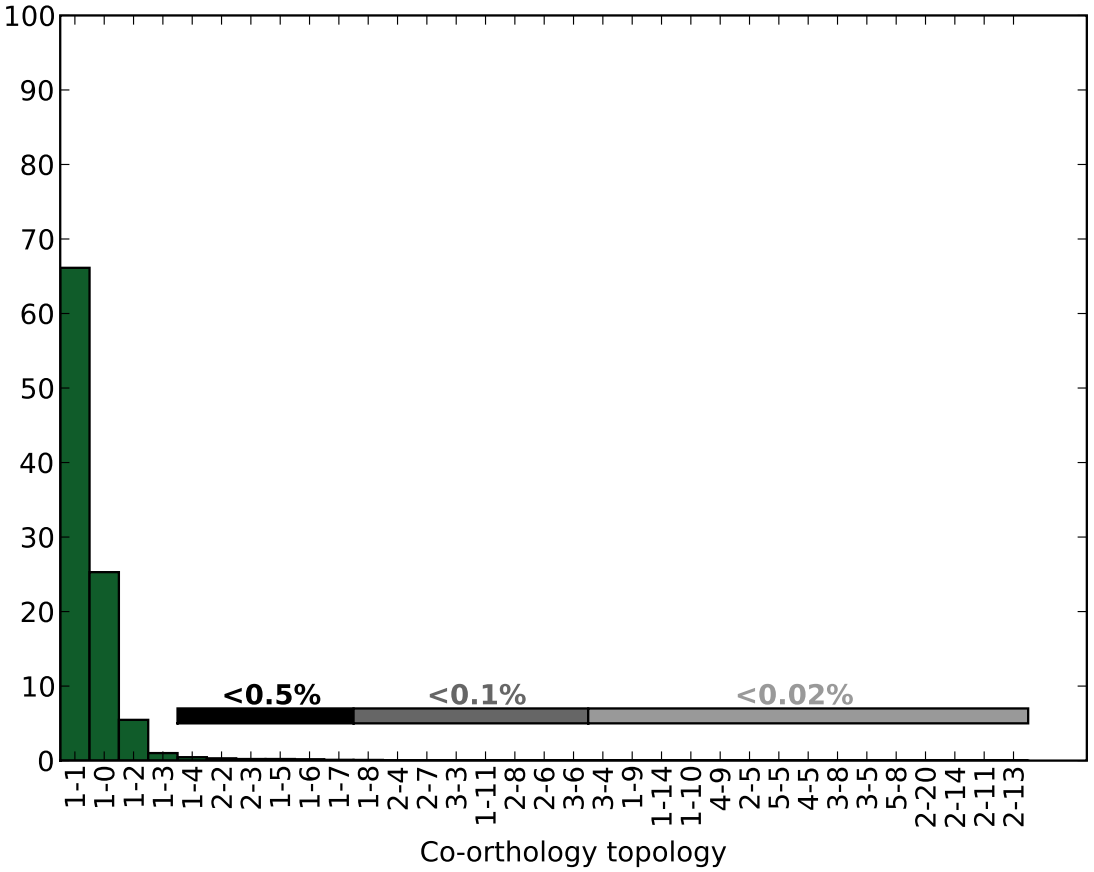


Figure 2.3: Distribution of OrthoMCL co-ortholog topologies for the five genomes described in Table 2.3. A co-ortholog topology describes the number of genes within an OrthoMCL cluster from the two species. For example, the topology 1-2 suggests a single gene in species A shares co-orthology with two genes in species B. It is not possible to know which of the two genes in species B is the true ortholog and which is an out-paralog. The topology 1-1 suggests a pair of orthologs from two species.

the database (Figure 2.2).

2.2.7 Models of orthology-based essentiality inference

Evolutionary pressures can cause genetic divergence between isolated strains of a species, resulting in the formation of new species, this process is termed "speciation". After speciation, genes that are essential for life must be maintained, and genes which offer a competitive advantage (fitness) to the new species may also be maintained. However, if the cost of expressing the gene into a protein is larger than the benefit of the function, the gene may eventually be lost from the genome. By considering these factors, it was possible to construct simple hypotheses and corresponding models to predict essentiality from the inferred phylogenic relationships observed between genomes. The hypotheses are detailed below and in Figure 2.4. Note that a genome that is being assessed for essentiality (the "predicted" genome), may have its phylogenetic inferred relationships analyzed against multiple species genomes (the "predictor" genome(s)).

2.2.7.1 Hypothesis 1

If a gene is conserved after a speciation event, then its function is required for survival and so the gene is essential. Model 1 essential rule: *The predicted gene is part of a co-orthologous group, it may be a single ortholog, or have multiple in-paralogs and out-paralogs.*

2.2.7.2 Hypothesis 2

If a gene is conserved after a speciation event, then its function is required for survival and is considered essential, however if the gene has subsequently been

duplicated there is potential redundancy of function and the gene is not individually essential. Model 2 essential rule: *The predicted gene has an ortholog or out-paralogs, but no in-paralogs.*

2.2.7.3 Hypothesis 3

A gene function is more probable to be essential, if the co-orthologous gene function in a related species is known to be essential. Model 3 essential rule: *The predicted gene satisfies the conditions of model 1 AND any one of the predictor co-orthologs has been experimentally verified as essential.*

2.2.7.4 Hypothesis 4

A gene function is more probable to be essential, if the co-orthologous gene function in a related species is known to be essential, however if the gene has subsequently been duplicated there is potential redundancy of function and the gene is not individually essential. Model 4 essential rule: *The predicted gene satisfies the conditions of model 2 AND any one of the predictor co-orthologs has been experimentally verified as essential.*

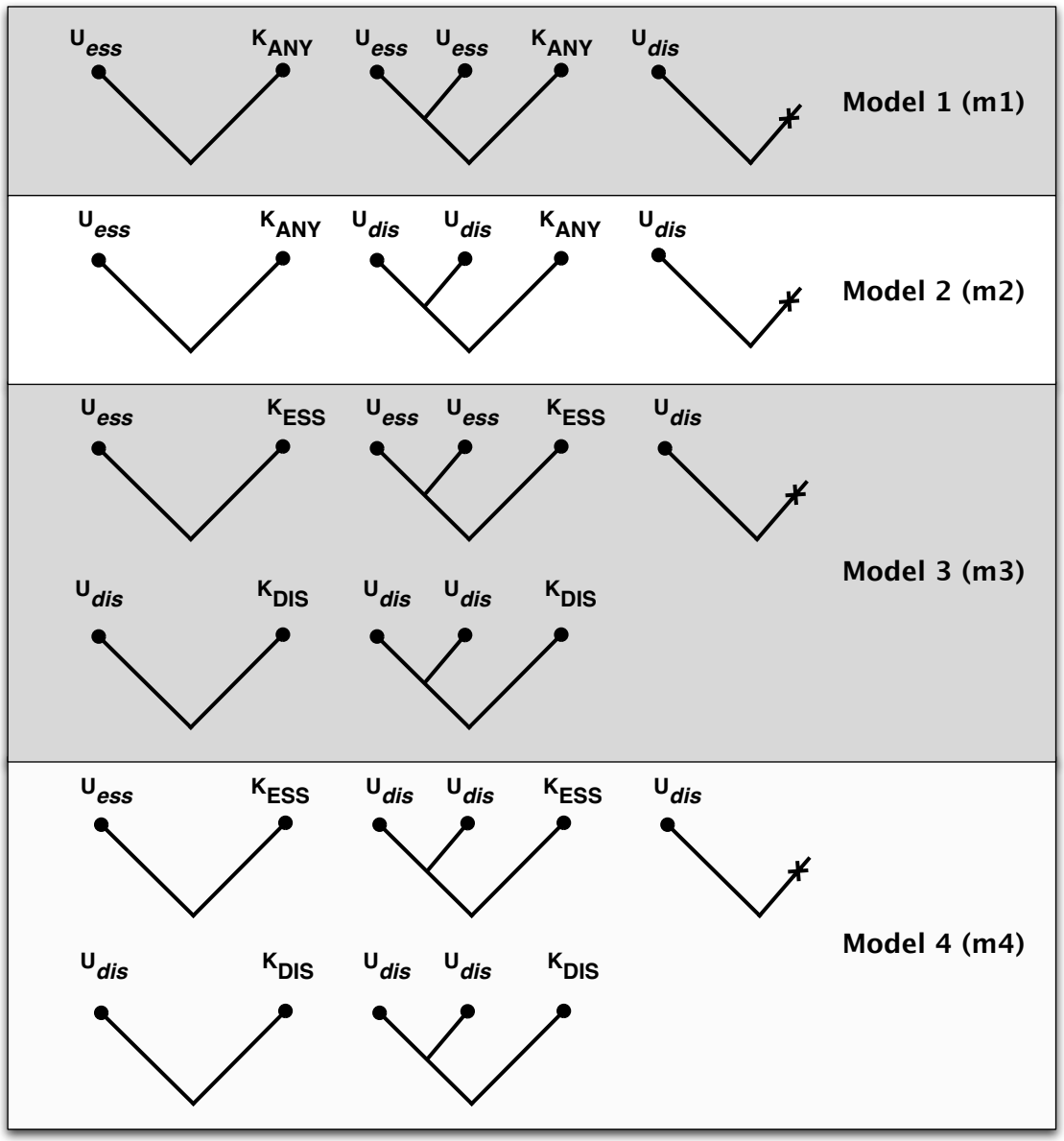
2.2.7.5 Hypothesis 5

A gene function is more probable to be essential, if the co-orthologous gene function in a related species is known to be essential, however if the gene has subsequently been duplicated there is potential redundancy of function and the gene is not individually essential. If the co-orthologous gene(s) from multiple genomes are known to be essential, then the chance of the predictor gene being essential is increased. Model 5 essential rule: *The predicted gene satisfies the conditions*

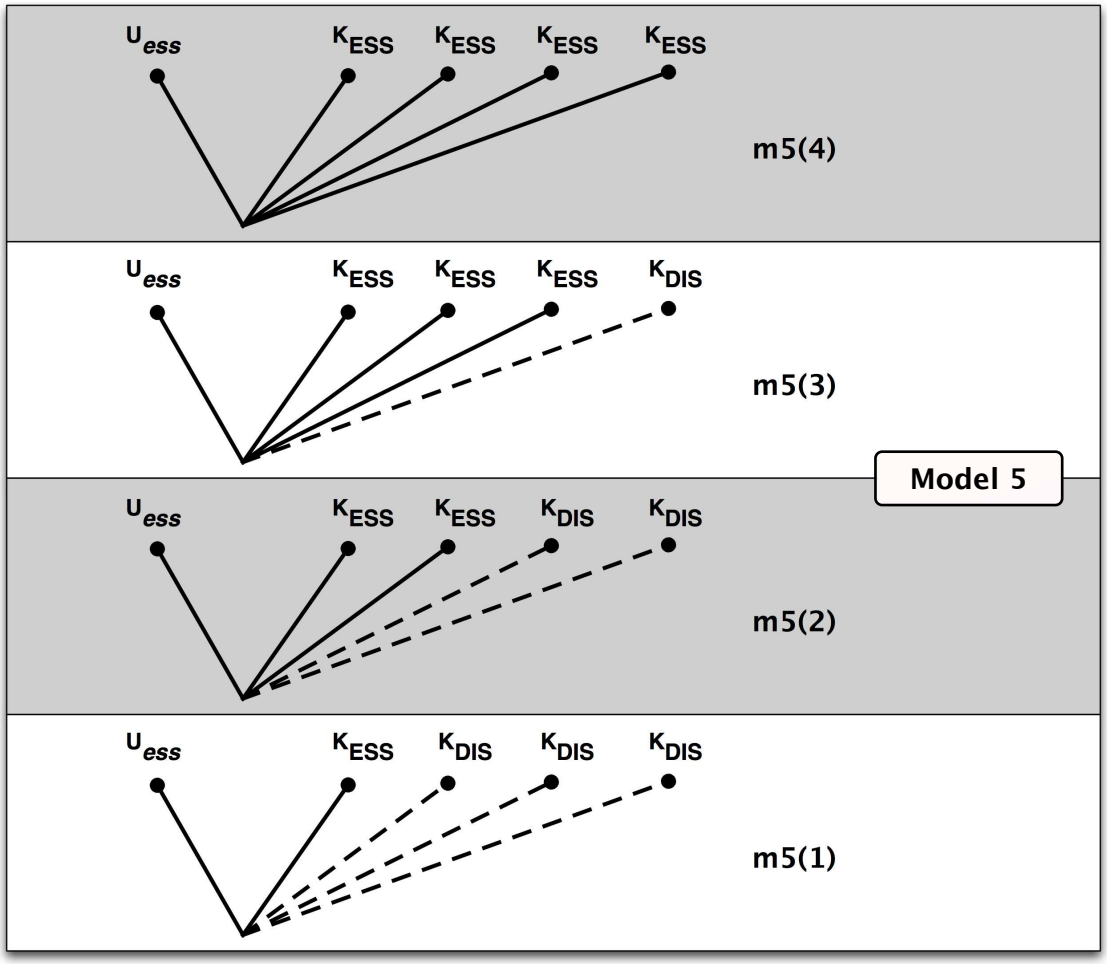
of model 4, and the number of times this is corroborated in other genomes, the greater the chance of essentiality. Note that in this case, results are denoted m5(N), where N is the number of species where the essentiality is corroborated.

Figure 2.4: Graphical representation of orthology based essentiality models. The four models of essentiality based on comparison with a single species (a), and the model based on multiple species (b). Each tree represents a speciation from a common ancestor. Each circle represents a gene from a species with experimentally known essentiality (**K**) or unknown essentiality (**U**). A crossed branch indicates an orthologous gene was not detectable or had been lost from the known species. The genes could be classified into observed essentials (**ESS**), observed dispensable/non-essentials (**DIS**), predicted essentials (***ess***), predicted dispensables (***dis***). A final gene classification (**ANY**) represented all genes from a species, where the model did not exploit observed essentiality data.

(a) Models 1 to 4. Each tree represents an outcome of an OrthoMCL orthology prediction, and how the model used the outcome to make an essentially prediction.



(b) Model 5, with four confidence levels. Dashed branches represent ortholog relationships that may exist but are not required for the prediction. Note that the tree is not a phylogenetic tree, but represents the speciations from a common ancestor.



2.2.8 Benchmark metric

When benchmarking the results of a binary classification experiment, it is important to consider both the sensitivity and specificity of the prediction method. Sensitivity or True Positive Rate (TPR) is a measure of how many correctly predicted positive outcomes are observed out of all the possible positive training

samples. Specificity or True Negative Rate (TNR) is a measure of how many incorrectly predicted positive outcomes are observed out of all possible negative training samples. In practical terms high sensitivity is not useful if specificity is poor. Often the most useful classifiers are those which provide a good balance of both, however in many cases, the preference for the balance is biased by the intended uses of the predictions. A Receiver Operating Characteristic (ROC) plot is a method for judging the effectiveness of altering classification parameters on this balance. By convention when plotted on a graph, the False Positive Rate (FPR) is on the X-axis and the TPR is on the Y-axis. The perfect classification system (with no false positives and no false negatives) would lie at the coordinate 0,1. Any predictions that lie on the line from coordinates [0,0] to [1,1] represent predictions no better than random guesses. A simple measure of the overall performance of a ROC point, is the Euclidean distance from the perfect classification point (PC_d), where a lower PD_d is preferable. In practise, when applying the predictive methods to a proteome, there will be a greater interest in the predicted essential proteins rather than the predicted non-essentials. The positive predictive value reflects how much confidence can be placed in positive (essential) predictions.

$$TPR = TP/(TP + FN) \tag{2.1}$$

$$TNR = TN/(FP + TN) \tag{2.2}$$

$$FPR = 1 - TNR \tag{2.3}$$

$$PPV = 100(TP/(TP + FP)) \tag{2.4}$$

$$PC_d = \sqrt{(0 - FPR)^2 + (1 - TPR)^2)} \tag{2.5}$$

where TP is the true positive count, TN is the true negative count, FP is the false positives count and FN is false negative count.

2.3 Results

2.3.1 Benchmark of homology inference

The performance of predicting essential genes by the inference of homology with known essential genes was tested using the benchmark described in Section 2.2.5. Each of the five proteomes of the benchmark species (Table 2.3), were compared against both the DEG database, and the essential proteins of the remaining four proteomes and where homology was inferred, those proteins were predicted essential. All predicted essential genes then compared to observed essentially status to measure accuracy. The performance of the essentiality inference, in terms of average sensitivity and specificity for all five proteomes is shown in Figure 2.5. A full breakdown of the performance of essentiality inference for each species is shown in appendix Table A.2. ROC plots of performance in each of the five species are shown in appendix Figure A.1.

The results suggest that with strict parameters (80-95% target coverage), increasing the size of the essential genes database increased the accuracy of predictions. With loose parameters (5-75% target coverage), the increase in the TPR

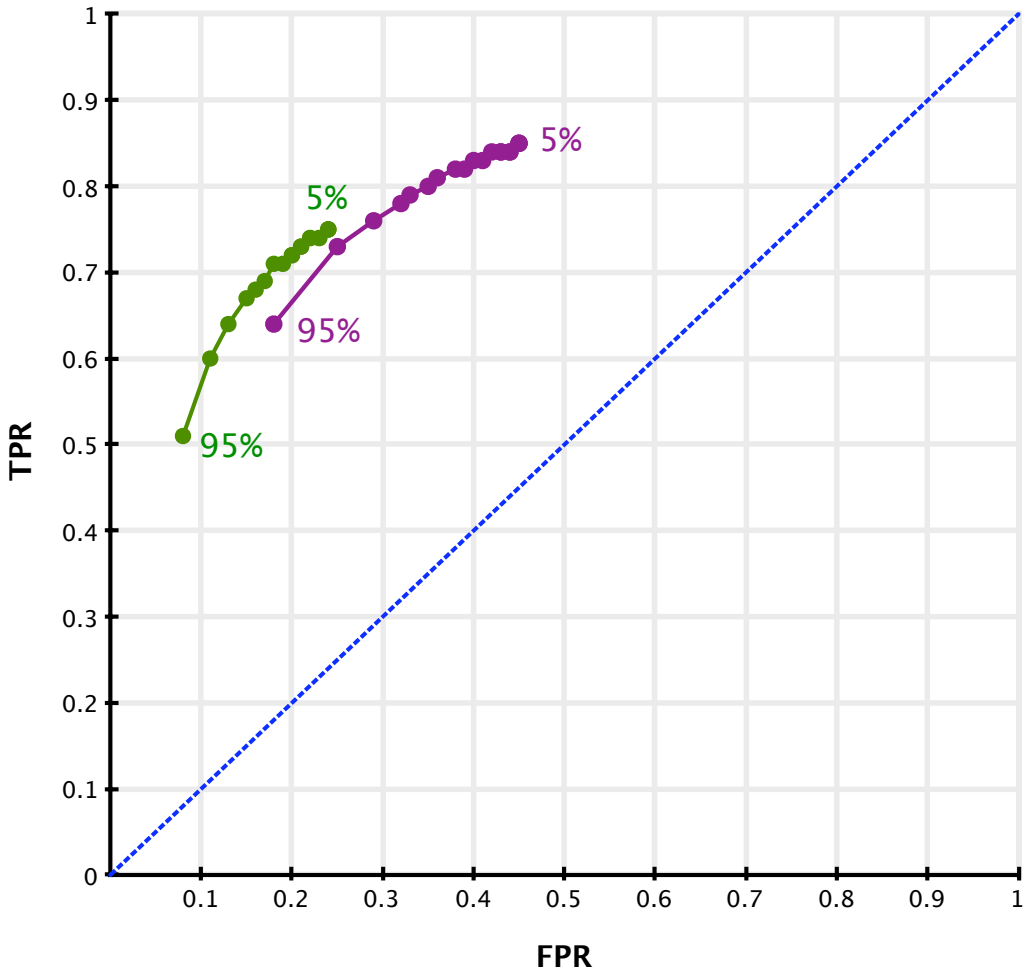


Figure 2.5: Benchmark of homology based essentiality prediction. Average performance of predicted essential proteins in five species using homology to two datasets of observed essential proteins. Essential datasets were: four predictor proteomes (green), and DEG (purple). Labels show the alignment coverage (%) of the observed essential sequence, required to infer homology. Individual performance of prediction for each species is shown in Appendix A.2.

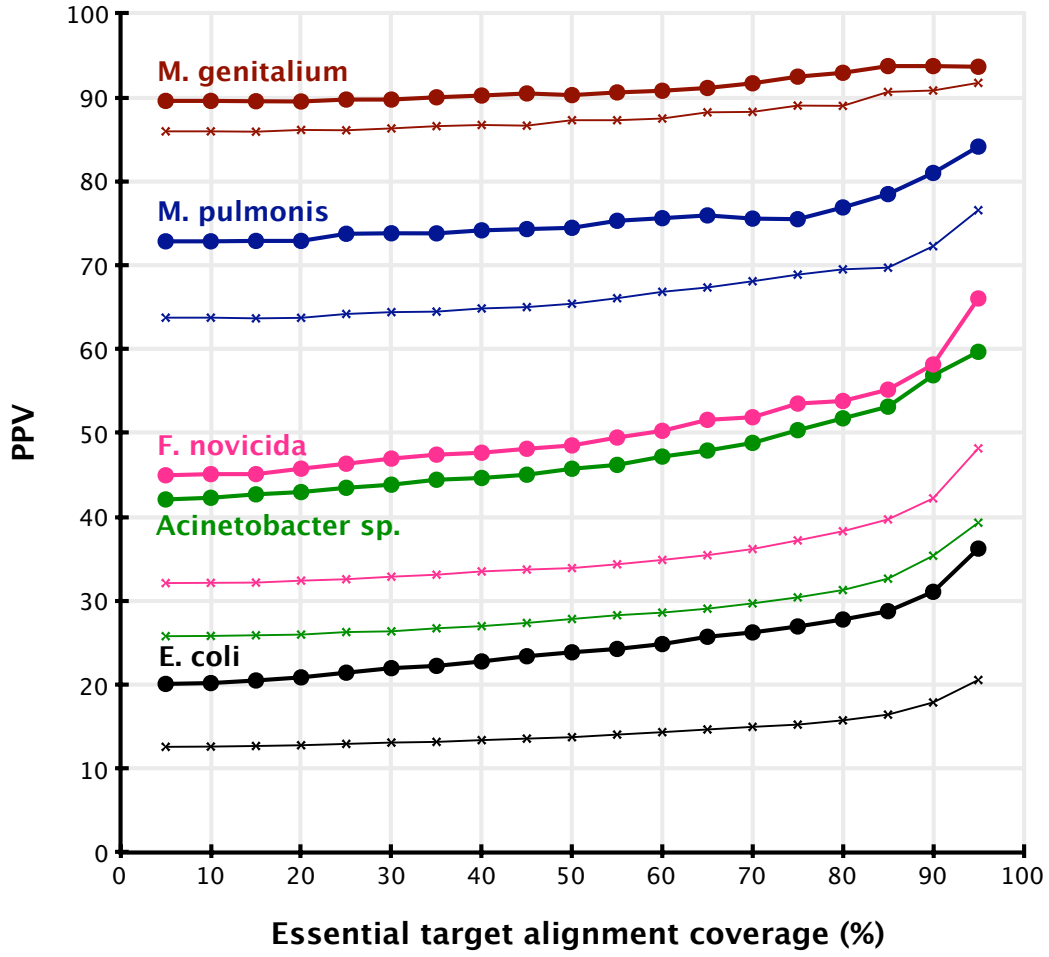


Figure 2.6: The positive predictive value (PPV) of the homology-based model of essentiality for each species in the benchmark. Against the four predictor proteome (circles) and against DEG (crosses). The results must be interpreted with respect to Table 2.3, as although predicted essential proteins in *M. genitalium* have the greatest chance of being essential (up to 94%), a proteins picked at random from this species would have a PPV of 80. Conversely, in *E. coli*, and random protein would be essential 7% of the time, but this chance can be increased to 36% using the most specific homology-based model.

is lower than the increase in the FPR. This suggests that as the number of species with experimentally observed essential genes increases, inference of essentiality by homology may become increasingly better at predicting true essentials from a proteome. However, using a larger essential proteins database also increases the overprediction (FPs) of essential proteins. A significant problem with this approach is that if a gene is correctly observed to be non-essential in multiple experiments, but incorrectly observed to be essential in just one experiment, this incorrect essential could "pollute" the essential database. As many proteins are members of large homologous families, this pollution could rapidly increase the overprediction of essentiality.

The accuracy of the inference of essentiality on the five species varied significantly (see Figure A.1). There may be multiple species specific factors that effect the performance, such as the size of genome, the evolutionary distance from species in the essential proteins database and the environmental niche the species occupies. Of the five species benchmarked here, the method performs the worst (in terms of sensitivity) on *M. genitalium*. Where other bacterial species may have functional redundancy within their proteomes, *M. genitalium* has a very small proteome (475 proteins) of which nearly 80% are essential. This lack of redundancy may mean many essential proteins of *M. genitalium* are specifically essential to itself, and the equivalent the functions in other species are performed by multiple not-individually-essential proteins.

If this was the only factor influencing the performance, then it would be expected that the essentials of the related species *M. pulmonis* would also be predicted poorly. However Figure 2.5 shows that *M. pulmonis* performs better against both essential databases. As *M. genitalium* essentials are present in both

databases, and as *M. genitalium* has more essential proteins than *M. pulmonis* it could be that *M. genitalium* is an ideal species to predict other species with, but a difficult species to be predicted itself.

Of the remaining three species, *E. coli* is consistently the best performer, outperforming *F. novicida* and *Acineobacter sp.* ADP1. Amongst other factors, the fact that the known *E. coli* essentials were verified using the "gold-standard" (gene-by-gene) method, then any predicted essentials were being validated against the most accurate data. As *Acineobacter sp.* ADP1 essentials was also verified using this gold-standard method, it could be expected that it would also perform well. Against the DEG database, its performance was comparable to that of *E. coli*, but against the four proteomes database, its TPR was much smaller. The main factor for this disparity is likely to be that the essentials of *Acineobacter sp.* ADP1 were observed on minimal medium, where the other four essential sets were observed on rich medium. This could result in those proteins involved in the biosynthesis of essential compounds being classified as non-essential thus reducing the TPR.

2.3.2 Benchmark of orthology inference

The five models of essentiality (described in Section 2.2.7) were applied to each genome-genome orthology comparison, and the predicted essential and non-essential genes correlated with the experimentally derived classifications. The results of the benchmark study are shown in Table 2.4, and in terms of sensitivity and specificity, in Figure 2.7.

It has been hypothesized, that genes conserved after speciation events are

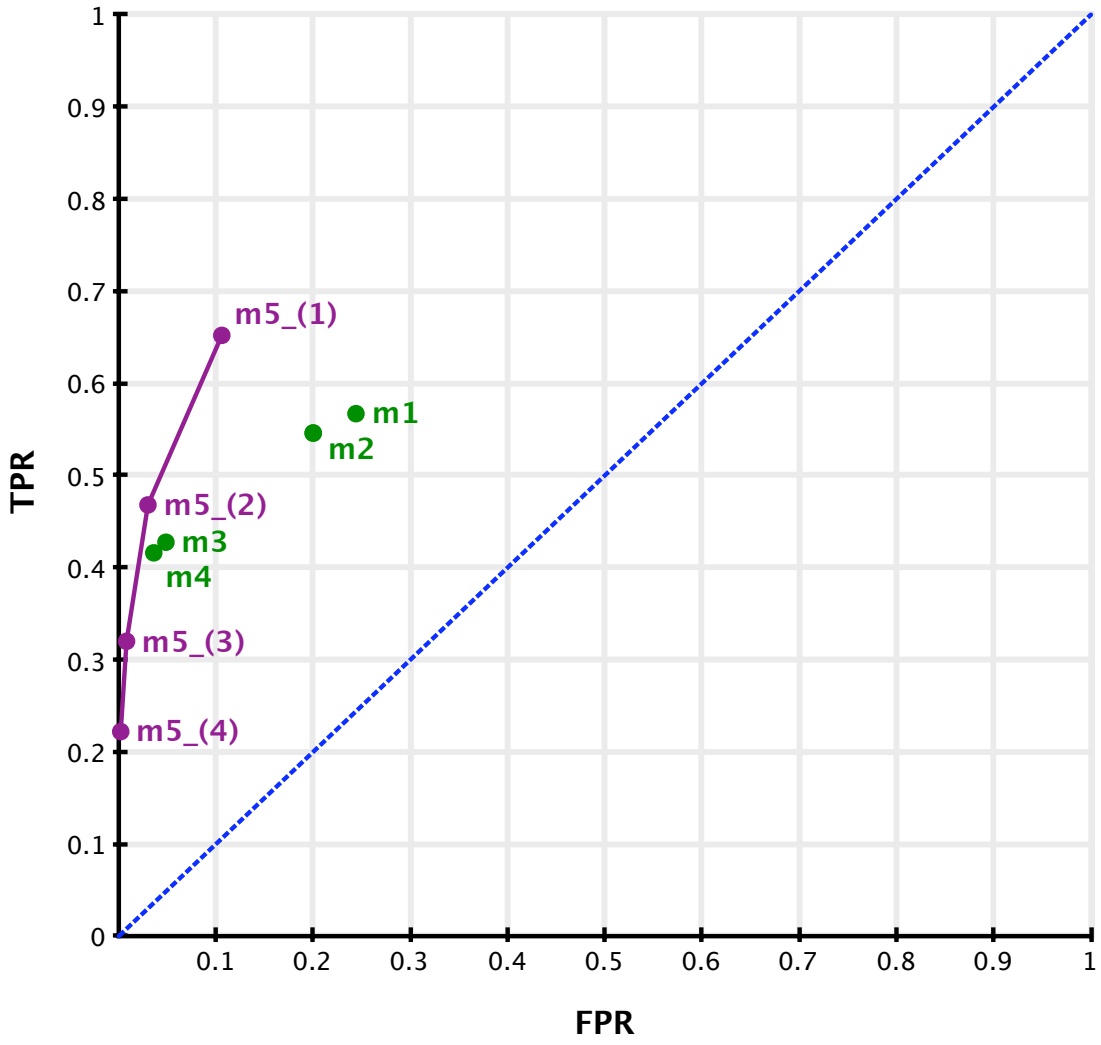


Figure 2.7: Benchmark of orthology based essentiality predictions. Data points represent the average values for all species in the benchmark. The models 1 to 4 in green. Model 5 in purple, the number in brackets indicates the minimum number of genomes which the predicted gene was required to share a known essential ortholog. Individual performance of prediction for each species is shown in Appendix A.3.

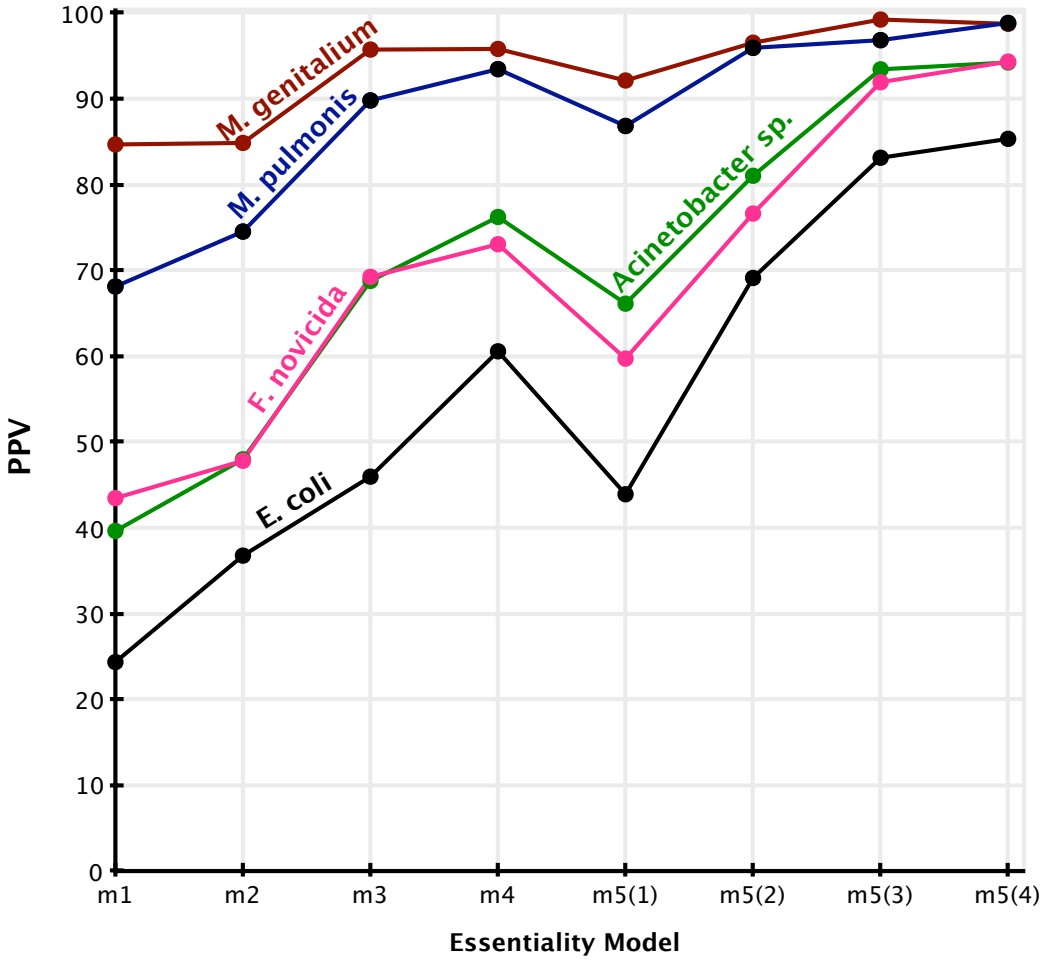


Figure 2.8: The positive predictive value (PPV) of each model of essentiality for each species in the orthology-based models.

more likely to be essential (Mushegian & Koonin, 1996). However, so far there has been little quantitative evidence to demonstrate these hypothesis. Here, it is shown using multiple species, that conservation of orthologs (model m1), is a positive predictor of essentiality. The extension of this hypothesis, that where these conserved orthologs have recently duplicated, to form in-paralogs, the chance of individual essentiality is reduced was assessed by model m2. If the hypothesis was correct, then model m2 should have performed better than model m1 at predicting essential genes. Across all species, on average, there was little difference in the overall performance of models m1 and m2 (PC_d of 0.50 and 0.50 respectively). However, the PPV of model m2 (Figure. 2.8) was always greater than model m1. It should be noted that the increase in PPV between models m1 and m2, has some correlation to proteome size, those species with very small proteomes such as *M. genitalium*, have far fewer in-paralogs and as such are less effected by the constrictions of model m2. Conversely, in the relatively large and redundant proteome of *E. coli*, in-paralogs appear to have duplicated the functions of a larger number of ancestrally essential proteins.

It has also been suggested, that genes conserved after speciation events are more likely to be essential, if an ortholog is known to be essential in another species (Aguero *et al.*, 2008). Here, it is shown using multiple species, that conservation of known essential orthologs (model m3), is a positive predictor of essentiality. Across all species, on average, there was little difference in the overall performance of models m3 and m4 (PC_d of 0.57 and 0.59 respectively). However, the PPV of of model m4 (Figure. 2.8) was always greater than model m3

Extending these models, to test if conserved essentiality in multiple comparison species (models m5(1)-m5(4)) improved the probability of essentiality. It

is clear from Figure 2.8, that the presence of an essential ortholog, in multiple species, greatly increases the chance of being essential. However, what is also apparent from Figure 2.4 is that with the greater number of species used to corroborate essentiality, while the PPV increases, the recall (TPR) of the predictions diminishes. In summary, as the models get more specific the enrichment of true essentials in the predicted set is increased. There is very high confidence that the set of essentials predicted with model m5(4) are truly essential, but the set will only be a small subset ($\approx$ 20%) of the entire essential set of the proteome.

Taking the specific example of *E. coli*, a previous essential prediction study (Holman *et al.*, 2009)(see 2.1.3.1) performed poorly on this species. In all the orthology based methods described here, *E. coli* was consistently the worst performer, in terms of PPV (Figure 2.8). By selecting a gene by chance in this species, it would be an essential gene 7% of the time. Using m1, and depending on the species used, the chance would be increased to between 15.4-31.2% (up to a 4.4 fold increase), by using m2 the chance increased to 18.8-53.8% (7.6 fold). This in some part explained the lack of performance in the Holman *et al.* (2009) method, which did not take account of paralogs. In models m3 and m4, the enrichment of essentials over a random guess for *E. coli* was up to 8-fold and 9-fold respectively. Using the most specific model (m5(4)), the proportion of true essentials in the *E. coli* predictions was 85.3%, or a 12-fold increase over chance. While the more specific models, performed well at enriching real essential genes in the predicted set, they did so at a cost of sensitivity, in the case of *E. coli*, only 81 of the known 296 essential genes (27%) were predicted in the m5(4) model.

Applying each of the benchmarked models consecutively gives a spectrum of predictions of varying confidence, thereby uniquely providing the means to

prioritize targets by the confidence in their likely essentiality in a quantitative manner.

2.3.3 The cost of specificity in the essentiality models.

The most specific model of essentiality for all five species benchmarked was model m5(4) (Table. 2.4). Depending on the species being examined, the resulting set of essential predictions were composed of between 80-99% observed essentials (Figure 2.8). However, with this specificity, the recall of essentials was poor and reduced the size of the essential sets to between 79-95 proteins. In total, across five species, 114 unique proteins were predicted as essential using model m5(4). These proteins are summarized in Appendix A.3. The proteins are mainly involved in functional classes fundamental across all cellular-organisms including, tRNA metabolism, DNA metabolism, protein synthesis, cell division, transcription and energy metabolism. Within this set there are many proteins that are also the targets of current drugs (*gyrA*, *gyrB*, *parE*, *rpoA*, *rpoB*, *rpoC*, *folA*, *ileS* and *fusA*)(Silver, 2011), and multiple protein subunits of the ribosome, of which both the protein and RNA components have been a common target for a diverse array of antibiotics (Yonath, 2005). While these prioritized targets have confidence of being essential, they are also limited in number, and many have previously been exploited for drug discovery. Drugs against these targets are also likely to be broad-spectrum, which is often useful, but can also facilitate the rapid rise of resistance.

Conversely, the most sensitive model of essentiality benchmarked was homology to proteins in the database of essential genes (DEG). This method also had

the advantage of being computationally quicker than the orthology based methods. The drawback of this approach is the lack of specificity, and the predicted essentials sets contain a smaller proportion of truly essential proteins.

Table 2.4: Orthology Benchmark Details for each species. (where **Pr.** = predicted species; **Use.** = predictor species; **P** = known essentials; **N** = known non-essentials; **Psize** = predicted species proteome size, **PPV** = positive predictor value = % essentials in predictions))

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m1	Acine.	E.coli	499	2808	15.1	399	1095	1713	100	3307	26.7	0.8	0.39		
m1	Acine.	Franc.	499	2808	15.1	338	606	2202	161	3307	35.8	0.68	0.22		
m1	Acine.	M.geni.	499	2808	15.1	129	131	2677	370	3307	49.6	0.26	0.05		
m1	Acine.	M.pulm.	499	2808	15.1	134	155	2653	365	3307	46.4	0.27	0.06		
m1	E.coli	Acine.	296	3853	7.1	244	1343	2510	52	4149	15.4	0.82	0.35		
m1	E.coli	Franc.	296	3853	7.1	229	828	3025	67	4149	21.7	0.77	0.21		
m1	E.coli	M.geni.	296	3853	7.1	106	234	3619	190	4149	31.2	0.36	0.06		
m1	E.coli	M.pulm.	296	3853	7.1	118	287	3566	178	4149	29.1	0.4	0.07		
m1	Franc.	Acine.	390	1329	22.7	302	606	723	88	1719	33.3	0.77	0.46		
m1	Franc.	E.coli	390	1329	22.7	306	653	676	84	1719	31.9	0.78	0.49		
m1	Franc.	M.geni.	390	1329	22.7	136	106	1223	254	1719	56.2	0.35	0.08		
m1	Franc.	M.pulm.	390	1329	22.7	142	129	1200	248	1719	52.4	0.36	0.1		
m1	M.geni.	Acine.	378	97	79.6	189	35	62	189	475	84.4	0.5	0.36		
m1	M.geni.	E.coli	378	97	79.6	200	36	61	178	475	84.7	0.53	0.37		
m1	M.geni.	Franc.	378	97	79.6	199	30	67	179	475	86.9	0.53	0.31		
m1	M.geni.	M.pulm.	378	97	79.6	237	50	47	141	475	82.6	0.63	0.52		

Continued on next page

Table 2.4 – *Continued from previous page*

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m1	M.pulm.	Acine.	310	472	39.6	179	83	389	131	782	68.3	0.58	0.18		
m1	M.pulm.	E.coli	310	472	39.6	192	117	355	118	782	62.1	0.62	0.25		
m1	M.pulm.	Franc.	310	472	39.6	189	87	385	121	782	68.5	0.61	0.18		
m1	M.pulm.	M.geni.	310	472	39.6	222	80	392	88	782	73.5	0.72	0.17	0.567	0.244
m2	Acine.	E.coli	499	2808	15.1	392	926	1882	107	3307	29.7	0.79	0.33		
m2	Acine.	Franc.	499	2808	15.1	335	468	2340	164	3307	41.7	0.67	0.17		
m2	Acine.	M.geni.	499	2808	15.1	119	75	2733	380	3307	61.3	0.24	0.03		
m2	Acine.	M.pulm.	499	2808	15.1	126	87	2721	373	3307	59.2	0.25	0.03		
m2	E.coli	Acine.	296	3853	7.1	241	1042	2811	55	4149	18.8	0.81	0.27		
m2	E.coli	Franc.	296	3853	7.1	223	609	3244	73	4149	26.8	0.75	0.16		
m2	E.coli	M.geni.	296	3853	7.1	100	86	3767	196	4149	53.8	0.34	0.02		
m2	E.coli	M.pulm.	296	3853	7.1	111	122	3731	185	4149	47.6	0.38	0.03		
m2	Franc.	Acine.	390	1329	22.7	297	522	807	93	1719	36.3	0.76	0.39		
m2	Franc.	E.coli	390	1329	22.7	301	584	745	89	1719	34.0	0.77	0.44		
m2	Franc.	M.geni.	390	1329	22.7	130	78	1251	260	1719	62.5	0.33	0.06		
m2	Franc.	M.pulm.	390	1329	22.7	135	96	1233	255	1719	58.4	0.35	0.07		
m2	M.geni.	Acine.	378	97	79.6	179	33	64	199	475	84.4	0.47	0.34		
m2	M.geni.	E.coli	378	97	79.6	191	34	63	187	475	84.9	0.51	0.35		

Continued on next page

Table 2.4 – *Continued from previous page*

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m2	M.geni.	Franc.	378	97	79.6	189	28	69	189	475	87.1	0.5	0.29		
m2	M.geni.	M.pulm.	378	97	79.6	233	48	49	145	475	82.9	0.62	0.49		
m2	M.pulm.	Acine.	310	472	39.6	167	54	418	143	782	75.6	0.54	0.11		
m2	M.pulm.	E.coli	310	472	39.6	180	84	388	130	782	68.2	0.58	0.18		
m2	M.pulm.	Franc.	310	472	39.6	175	57	415	135	782	75.4	0.56	0.12		
m2	M.pulm.	M.geni.	310	472	39.6	216	58	414	94	782	78.8	0.7	0.12	0.546	0.2
m3	Acine.	E.coli	499	2808	15.1	205	40	2768	294	3307	83.7	0.41	0.01		
m3	Acine.	Franc.	499	2808	15.1	225	86	2722	274	3307	72.3	0.45	0.03		
m3	Acine.	M.geni.	499	2808	15.1	123	99	2709	376	3307	55.4	0.25	0.04		
m3	Acine.	M.pulm.	499	2808	15.1	126	72	2736	373	3307	63.6	0.25	0.03		
m3	E.coli	Acine.	296	3853	7.1	205	224	3629	91	4149	47.8	0.69	0.06		
m3	E.coli	Franc.	296	3853	7.1	188	143	3710	108	4149	56.8	0.64	0.04		
m3	E.coli	M.geni.	296	3853	7.1	103	191	3662	193	4149	35.0	0.35	0.05		
m3	E.coli	M.pulm.	296	3853	7.1	111	140	3713	185	4149	44.2	0.38	0.04		
m3	Franc.	Acine.	390	1329	22.7	225	122	1207	165	1719	64.8	0.58	0.09		
m3	Franc.	E.coli	390	1329	22.7	188	43	1286	202	1719	81.4	0.48	0.03		
m3	Franc.	M.geni.	390	1329	22.7	131	74	1255	259	1719	63.9	0.34	0.06		
m3	Franc.	M.pulm.	390	1329	22.7	131	65	1264	259	1719	66.8	0.34	0.05		

Continued on next page

Table 2.4 – *Continued from previous page*

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m3	M.geni.	Acine.	378	97	79.6	126	6	91	252	475	95.5	0.33	0.06		
m3	M.geni.	E.coli	378	97	79.6	105	3	94	273	475	97.2	0.28	0.03		
m3	M.geni.	Franc.	378	97	79.6	133	5	92	245	475	96.4	0.35	0.05		
m3	M.geni.	M.pulm.	378	97	79.6	209	14	83	169	475	93.7	0.55	0.14		
m3	M.pulm.	Acine.	310	472	39.6	128	10	462	182	782	92.8	0.41	0.02		
m3	M.pulm.	E.coli	310	472	39.6	114	9	463	196	782	92.7	0.37	0.02		
m3	M.pulm.	Franc.	310	472	39.6	132	17	455	178	782	88.6	0.43	0.04		
m3	M.pulm.	M.geni.	310	472	39.6	209	37	435	101	782	85.0	0.67	0.08	0.428	0.049
m4	Acine.	E.coli	499	2808	15.1	204	39	2769	295	3307	84.0	0.41	0.01		
m4	Acine.	Franc.	499	2808	15.1	225	69	2739	274	3307	76.5	0.45	0.02		
m4	Acine.	M.geni.	499	2808	15.1	115	49	2759	384	3307	70.1	0.23	0.02		
m4	Acine.	M.pulm.	499	2808	15.1	118	41	2767	381	3307	74.2	0.24	0.01		
m4	E.coli	Acine.	296	3853	7.1	203	179	3674	93	4149	53.1	0.69	0.05		
m4	E.coli	Franc.	296	3853	7.1	186	107	3746	110	4149	63.5	0.63	0.03		
m4	E.coli	M.geni.	296	3853	7.1	97	61	3792	199	4149	61.4	0.33	0.02		
m4	E.coli	M.pulm.	296	3853	7.1	106	59	3794	190	4149	64.2	0.36	0.02		
m4	Franc.	Acine.	390	1329	22.7	224	107	1222	166	1719	67.7	0.57	0.08		
m4	Franc.	E.coli	390	1329	22.7	187	40	1289	203	1719	82.4	0.48	0.03		

Continued on next page

Table 2.4 – *Continued from previous page*

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m4	Franc.	M.geni.	390	1329	22.7	126	56	1273	264	1719	69.2	0.32	0.04		
m4	Franc.	M.pulm.	390	1329	22.7	126	47	1282	264	1719	72.8	0.32	0.04		
m4	M.geni.	Acine.	378	97	79.6	121	6	91	257	475	95.3	0.32	0.06		
m4	M.geni.	E.coli	378	97	79.6	100	3	94	278	475	97.1	0.26	0.03		
m4	M.geni.	Franc.	378	97	79.6	128	5	92	250	475	96.2	0.34	0.05		
m4	M.geni.	M.pulm.	378	97	79.6	207	12	85	171	475	94.5	0.55	0.12		
m4	M.pulm.	Acine.	310	472	39.6	123	7	465	187	782	94.6	0.4	0.01		
m4	M.pulm.	E.coli	310	472	39.6	108	5	467	202	782	95.6	0.35	0.01		
m4	M.pulm.	Franc.	310	472	39.6	126	8	464	184	782	94.0	0.41	0.02		
m4	M.pulm.	M.geni.	310	472	39.6	204	24	448	106	782	89.5	0.66	0.05	0.416	0.036
m5(1)	Acine.	all	499	2808	15.1	271	139	2669	228	3307	66.1	0.54	0.05		
m5(1)	E.coli	all	296	3853	7.1	227	290	3563	69	4149	43.9	0.77	0.08		
m5(1)	Franc.	all	390	1329	22.7	256	173	1156	134	1719	59.7	0.66	0.13		
m5(1)	M.geni.	all	378	97	79.6	223	19	78	155	475	92.1	0.59	0.2		
m5(1)	M.pulm.	all	310	472	39.6	217	33	439	93	782	86.8	0.7	0.07	0.652	0.106
m5(2)	Acine.	all	499	2808	15.1	196	46	2762	303	3307	81.0	0.39	0.02		
m5(2)	E.coli	all	296	3853	7.1	181	81	3772	115	4149	69.1	0.61	0.02		
m5(2)	Franc.	all	390	1329	22.7	200	61	1268	190	1719	76.6	0.51	0.05		

Continued on next page

Table 2.4 – *Continued from previous page*

Mod.	Pr.	Use.	P	N	% ess.	TP	FP	TN	FN	Psize	PPV	TPR	FPR	AvTPR	AvFPR
m5(2)	M.geni.	all	378	97	79.6	138	5	92	240	475	96.5	0.37	0.05		
m5(2)	M.pulm.	all	310	472	39.6	142	6	466	168	782	95.9	0.46	0.01	0.468	0.03
m5(3)	Acine.	all	499	2808	15.1	114	8	2800	385	3307	93.4	0.23	0		
m5(3)	E.coli	all	296	3853	7.1	103	21	3832	193	4149	83.1	0.35	0.01		
m5(3)	Franc.	all	390	1329	22.7	124	11	1318	266	1719	91.9	0.32	0.01		
m5(3)	M.geni.	all	378	97	79.6	117	1	96	261	475	99.2	0.31	0.01		
m5(3)	M.pulm.	all	310	472	39.6	121	4	468	189	782	96.8	0.39	0.01	0.32	0.008
m5(4)	Acine.	all	499	2808	15.1	81	5	2803	418	3307	94.2	0.16	0		
m5(4)	E.coli	all	296	3853	7.1	81	14	3839	215	4149	85.3	0.27	0		
m5(4)	Franc.	all	390	1329	22.7	83	5	1324	307	1719	94.3	0.21	0		
m5(4)	M.geni.	all	378	97	79.6	78	1	96	300	475	98.7	0.21	0.01		
m5(4)	M.pulm.	all	310	472	39.6	81	1	471	229	782	98.8	0.26	0	0.222	0.002

2.4 Conclusions and future direction

Phylogenetic approaches to predicting essential targets carry the advantage that they can be applied to genomes lacking experimental essentiality data. These approaches can be based purely on inferred phylogenetic information (e.g. the absence of paralogs within a genome, the presence of orthologs between genomes), or include essentiality data from related species (e.g. the presence of essential orthologs or homologs) or indeed a combination of the above.

This analysis was performed on prokaryote species, but in a study by Doyle *et al.* (2010) the application of similar methods to four eukaryotic species (*Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Mus musculus*) was addressed. As in prokaryotes, the presence of an ortholog (model m1) was always a positive predictor of essentiality. In the prokaryotes used here, the absence of paralogs improved the predictive power further, which was also the trend in eukaryotes with the exception of *M. musculus*, where the predictive power deteriorated. In both prokaryote and eukaryotes, the presence of an observed essential ortholog substantially increased the likelihood of essentiality, but at a cost to the recall of known essential genes.

As new sources of essentiality data become available, these can be added to the system to increase the range and sensitivity of the predictions further. The models described above have the advantage of having virtually no cost, in terms of man-hours, experimental overheads and CPU time.

The breadth of observed essential genomes has increased since this analyses, with the recent publications of the essential genomes of several bacterial including, the opportunistic pathogen *Streptococcus sanguinis*, the pathogen *Porphyromonas*

gingivalis and the model organism *Caulobacter crescentus* (Christen *et al.*, 2011; Klein *et al.*, 2012; Xu *et al.*, 2011). While the rate of new whole genome essentiality screens appears to be increasing, it is not at such a rate to have any significant effect on the computational overhead of the orthology methods.

When assessing a new pathogen for a drug discovery program defining the perturbative targets is a priority (Frearson *et al.*, 2007). By using the most prescriptive model of prediction (i.e. m5(4)), clearly the confidence of your essential set is high. However, given the small proportion of essential genes usually observed in a bacterial genome, and the low sensitivity of the method, the resulting set of essential genes may be very small. When these targets are subjected to other criteria such as druggability, the set will diminish further. Therefore, instead of using an individual model of essentiality, it is preferable to use all models, and rank the predictions by the selectivity of the model, which prioritizes the most confident essentials, but does not limit the range of targets available for consideration. A further prioritization could be achieved by pre-ranking the observed essential genomes with a weighting based on the quality of the experimental method employed. Quality is a subjective measure, but for the purposes of perturbation of a pathogen, those experiments which more precisely mimic the environment of a host are preferable (e.g. rich medium essentials rather than minimal medium). Other factors that effect the quality of predictions include the experimental method, random transposon mutagenesis has the tendency to over-predict essential genes, and as such prediction based solely on these sets could be weighted lower than those of gene-by-gene deletion sets. As the number of essential genomes increases, a bias in prediction ranking could be introduced if multiple evolutionary closely related species are used simultaneously.

Recently, Deng *et al.* (2011) published a machine learning method which amongst other factors, utilized intrinsic, calculable features of a protein (such as codon bias, hydrophobicity score, aromaticity and cellular location) to predict essentiality. In the future, new models such as these, or improved orthology models based on more experimentally derived essentiality sets, could be simply added to the benchmark, to further increase the spectrum of predictions.

Chapter 3

Application of phylogenomic inference of essentiality

3.1 Application of essentiality predictions to target prioritization in *Pseudomonas aeruginosa*

Pseudomonas aeruginosa is an important Gram-negative bacterial pathogen that is of major clinical significance as a cause of pneumonia, septic shock, urinary tract and gastrointestinal infections and a particular problem for Cystic Fibrosis patients and burn victims (Balcht & Smith, 1994; Kerr & Snelling, 2009). The tendency to form biofilms (Høiby *et al.*, 2001), the low permeability of the bacterial cellular envelope, the presence of multidrug efflux pumps and chromosomally-encoded antibiotic resistance genes all combine to make *P. aeruginosa* an intrinsically challenging pathogen. The challenge is exacerbated by the capacity of *P. aeruginosa* to acquire antibiotic resistance, either by genomic mutation or horizontal gene transfer of antibiotic resistance determinants. For these reasons the Aeropath project (http://www.aeropath.eu/) was undertaken to identify novel drug targets in *P. aeruginosa*. The Aeropath Target Database aids the identification of potential drug targets from the genome of *P. aeruginosa* by coupling a chemistry-led approach to predicting target druggability with information on gene essentiality, virulence factors, predicted selectivity, related bacterial orthologs and assessment of structural biology accessibility. In the database, perturbative targets are identified as either experimentally observed essential targets, phylogenomically inferred essential targets or as known virulence factors. The Aeropath Target Database was primarily developed by Dr. Richard Bickerton in the Hopkins group. My role in this project was to implement the essentiality

inference module of the database using the methods described in Chapter 2. The Aeropath Target Database is an Oracle relational database hosted locally at the University of Dundee. A web front end for accessing the database is available at http://aeropath.lifesci.dundee.ac.uk.

3.1.1 Motivation

In the case of *P. aeruginosa* there are two sets of published large-scale transposon knockouts (Jacobs *et al.*, 2003; Liberati *et al.*, 2006), identifying essential genes. In the method, those genes that were disrupted by the transposon and still viable were considered non-essential, and those genes that were never observed containing a transposon were considered essential. In some cases, the genes were only recovered with a single transposon insertion site close to either the 5′ or 3′-end, which could indicate that they were essential, but the transposon did not disrupt the essential region of the protein, those genes were termed "potentially essential". With this method, genes that are either small or located in transposition cold spots could be classified as essential by chance.

The Jacobs *et al.* (2003) mutagenesis study was performed on the PAO1 strain of *P. aeruginosa* and identified 773 genes as being essential (or 13.9% of the proteome) including 97 potentially essential genes. The Liberati *et al.* (2006) study was performed on the PA14 strain and the essential gene candidates observed that were also observed in the PA01 study were reported as essential, suggesting that 364 genes are essential (or 6.5% of the proteome). Therefore the Jacobs *et al.* (2003) set subsumes the Liberati *et al.* (2006) set. Essentiality depends on the context of the experimental conditions, and the ability of the experiment

to accurately assess each individual gene for essentiality. For these reasons, the experimentally derived set was augmented with predicted essential genes.

3.1.2 Essential gene prediction

Essentiality prediction was performed on the reference *P. aeruginosa* strain obtained from http://www.pseudomonas.com/. Essentiality prediction was performed using model m5(1) (see section 2.2.7.5), using all five of the benchmark species reported in Table 2.3 (page 32).

3.1.3 Results and discussion

The number of predicted essential genes in *P. aeruginosa* using model m5(1) was 572 (or 10.3% of the proteome). The overlap of the predicted set and the two experimentally derived sets is shown in Figure 3.1. Each source of perturbative (1.3.3) targets provides a different but overlapping set of proteins, the intersects of these sets provide additional levels of confidence of essentiality. In combination they suggest that 1,050 are potentially perturbative (or 18% of the proteome).

If the Liberati *et al.* (2006) set (364 genes) is taken as the true complement of essentials then the predictions produce a true positive rate (TPR) of 0.59, which is within the range of performance observed in the benchmark of model m5(1) (see Table 2.4). Conversely, if the Jacobs *et al.* (2003) set (773 genes) is taken as the true complement of essentials then the predictions produce a TPR of 0.39, which is comparatively poor compared to the benchmark. This difference in performance may be explained by the Jacobs *et al.* (2003) set being an overestimation of essential genes due to limitation in the experimental method, or that

the proteome size of *P. aeruginosa* (5,677) is larger than all of the predictor species (see Table 2.3). Where experimentally observed essential genes are available for a species, the predictive methods provide a means to prioritize essential genes from an over-estimated set, and supplement the observed set with essential genes which have been missed due to experimental limitations.

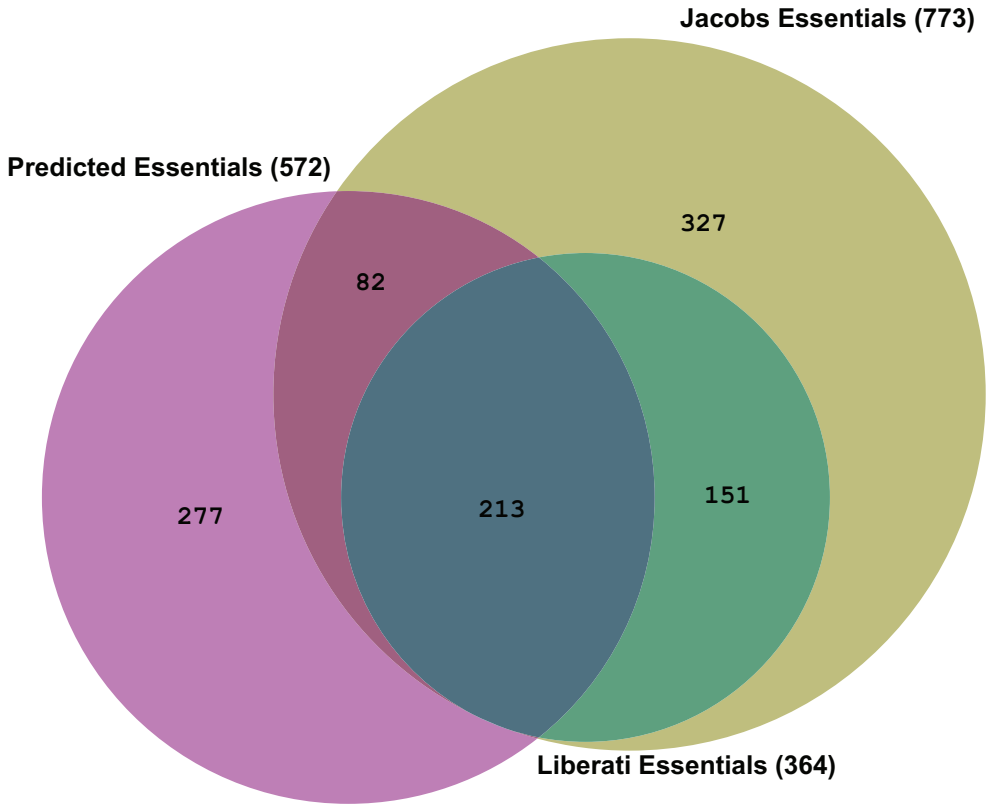


Figure 3.1: The overlap of the predicted essential genes in *P. aeruginosa*, with the experimentally observed essential genes of strain PAO1 (Jacobs *et al.*, 2003) and strain PA14 (Liberati *et al.*, 2006). Figure produced using BioVenn (Hulsen *et al.*, 2008).

3.2 Application of essentiality predictions to target prioritization in kinetoplastids

Our close collaborators in the Drug Discovery Unit (DDU) at the University of Dundee have a longstanding interest in kinetoplastid diseases. Target selection remains a crucial decision point in the drug discovery process and a lack of high quality validated targets has become a bottleneck for progress in the search for new antitrypanosomal therapeutics (Wyatt *et al.*, 2011). As such, having seen the example of the Aeropath Target Database, the DDU were keen to use a similar systematic approach to identify potential targets in the kinetoplastids. The result of this work was the Kinetoplastid Target Database (KTD)(`http://rapid.lifesci.dundee.ac.uk/KTD/`) which was built with the same technology as the Aeropath Target Database, but applied to multiple genomes. As before, my role was to supply the essentiality inference module. The KTD includes the genomes of *Trypanosoma brucei* (the causative agent of Human African trypanosomiasis or sleeping sickness in humans and nagana in animals), *Trypanosoma cruzi* (El-Sayed *et al.*, 2005) (the causative agent of Chagas disease), *Leishmania braziliensis* (Peacock *et al.*, 2007), *Leishmania infantum* (Peacock *et al.*, 2007) and *Leishmania major* (Ivens *et al.*, 2005)(the causative agents of Leishmaniasis). In total 7 different proteomes were included as three different genetic variants of *T. brucei* were covered: *T. brucei* strain TREU 927 (Berriman *et al.*, 2005), *T. brucei* Lister strain 427 (Becker *et al.*, 2004) and *T. brucei gambiense* (Jackson *et al.*, 2010).

3.2.1 Essentiality inference

The essentiality inference work capitalized on the genome scale RNAi knock-down experiments performed on *T. brucei* by Alsford *et al.* (2011). *T. brucei* is transmitted to mammalian hosts by the tsetse fly, and is subject to complex morphological changes that are induced as it switches between insect and mammal hosts, and during its life-cycle. The RNAi experiments were performed in 4 different induced samples, bloodstream-form cells grown for three (BF^{D3}) or six days (BF^{D6}), procyclic-form cells (PF) and differentiated cells (DIF). In this work *T. brucei* genes were considered essential if they were essential in the BF^{D3} and BF^{D6} forms, as it is in these forms that the trypanosomes multiply in the host and cause irreversible damage. The procyclic-form occurs exclusively in infection of the tsetse fly. The differentiated-cells set represented cells grown as bloodstream forms, induced into the non-dividing form that serve to re-infect tsetse flies (Brun *et al.*, 2010), and then grown as procyclic forms. While the PF and DIF essentials were not considered useful for essential drug-target prediction, the information was retained and could be utilized by KTD users if required. The phylogenomic models used for essentiality inference were Model 3 (has an essential ortholog) and Model 4 (has essential ortholog and no in-paralogs). All proteomes were obtained from TriTrypDB version 3.1 (Aslett *et al.*, 2010) or UniProt complete proteomes (16th June 2011) (Wu *et al.*, 2006). The essential protein sets were taken directly from the supplemental material of Alsford *et al.* (2011), using the cutoffs described by the authors to determine significant loss-of-fitness genes (i.e. significant and positive Z-score). Of the ≈7500 non-redundant protein-coding regions of the *T. brucei* genome, Alsford *et al.* (2011) obtained data for 7435 of

the regions. Of the protein-coding regions covered, 1908 (25.7%) were essential for fitness in the BF^{D3} set, and 2724 (36.6%) in the BF^{D6} set.

3.2.2 Results

Table 3.1 shows the proportions of each of the kinetoplastid proteomes that are predicted essential using the two models, and the effect of predicting based on permutations of the BF^{D3} and BF^{D6} essential proteins. *T. cruzi* Esmeraldo-Like is consistently predicted to have a smaller proportion of essential proteins than the other kinetoplasids, this is not only due to its larger genome size, as the number of proteins predicted essential are also consistently smaller (see Appendix Table A.4), and the cause of this is due to a smaller number of ortholog relationships inferred by orthomcl (see Appendix Table A.5). The remaining kinetoplastids are predicted to have similar proportions of essential proteins in each set, which is unsurprising given the similar life-cycles of these species. Its is important to note, that in the three *Leishmania* species, only $\approx 70\%$ of proteins have any co-ortholog to the reference *T. brucei* strain, and so 30% of these proteomes cannot be predicted as essential or non-essential by this method. By using the $BF^{D3} \cap BF^{D6}$:m4 results, the user can select those targets that are likely to be essential for fitness throughout the host-infection phase, and are likely to be individually essential. However, this would limit the potential targets to $< 15\%$ of the proteomes. Larger sets can be selected by considering potential polypharmacology targets using m4 or by considering targets only essential in one of the host-infection phases (e.g. $BF^{D3} \cup BF^{D6}$:m4).

Currently, the DDU is utilizing these essentiality predictions along with other

information available in the KTD to prioritize potential drug targets for validation studies.

		$\mathbf{BF^{D3} \cup BF^{D6}}$		$\mathbf{BF^{D6}}$		$\mathbf{BF^{D3}}$		$\mathbf{BF^{D3} \cap BF^{D6}}$	
kinetoplastid	**Psize**	m3	**m4**	m3	**m4**	m3	**m4**	m3	**m4**
T. cruzi Esmeraldo-Like	10342	25.0	**21.5**	22.1	**18.8**	15.6	**12.8**	12.7	**10.1**
T. brucei gambiense	9668	31.8	**29.3**	28.0	**25.6**	19.5	**17.7**	15.7	**14.0**
T. brucei Lister strain 427	8529	38.2	**33.3**	33.6	**29.1**	23.8	**19.9**	19.3	**15.7**
L. infantum	8033	31.5	**29.2**	27.9	**25.7**	19.6	**17.8**	16.0	**14.3**
L. major	8045	31.2	**29.2**	27.6	**25.8**	19.4	**17.8**	15.8	**14.3**
L. braziliensis	7809	31.3	**29.0**	27.8	**25.6**	19.6	**17.7**	16.0	**14.3**

Table 3.1: Percentage of predicted essential proteins in 6 kinetoplastids. (where **Psize** = proteome size; BF^{D3}/BF^{D6} = bloodstream form after 3/6 days respectively). The essential proteins of *T. bruci* strain TREU 927. were used to infer essentiality using models m3 and m4 (as described in Chapter 2.2.7). The essential predictions were based on proteins shown to be essential in either BF^{D6}, BF^{D3}, both forms or either form.

3.3 Application of essentiality predictions to target prioritization in *Schistosoma mansoni*

S. mansoni is a trematode flatworm and one of the causative agents schistosomiasis. Schistosomiasis is a significant cause of morbidity in tropical regions, where an estimated 600 Million people are at a significant risk of infection. Currently the only treatment available is praziquantel, a drug that has been in use over 20 years, and there are fears that resistance is developing (Doenhoff *et al.*, 2009).

3.3.1 Motivation

Dr. Quentin Bickle's group from the London School of Hygiene and Tropical Medicine (LSHTM), aimed to develop new targets for schistosomiasis interven-

tion. At the start of this project they were in the early stages of setting up facilities to screen *S. mansoni* targets with potential small-molecule inhibitors, and established RNA interference (RNAi) techniques to validate gene targets for essentiality. The two requirements for the project were targets that were potentially essential, and targets that had known chemical inhibitors validated against them, or against homologs, as a starting point for compound development.

3.3.2 Essential gene prediction

The organism used to infer essentiality was *Caenorhabditis elegans*. The genome of *C. elegans* was available (Hillier *et al.*, 2005), and downloaded from wormbase version 205 (Rogers *et al.*, 2008) that contained ≈24k proteins. A systematic functional analysis of the *C. elegans* genome, using RNAi, was available (Kamath *et al.*, 2003). Kamath *et al.* (2003) assigned each gene to a functional class based on the phenotype, and the functional classes used to define "essential" genes here were:

- **Embryonic lethality (Emb)**, defined as >10% dead embryos.
- **Sterile (Ste)**, required a brood size of <10 (**wild-type** worms under similar conditions typically have >100 progeny).
- **Sterile progeny (Stp)**, progeny brood size of <10.

Using these functional classes produced only 1170 essential genes in the *C. elegans* set. As only one essential genome was considered appropriate, the most specific essentiality model available was m4 (see section 2.2.7.4). The proteome of *S. mansoni* was download from Sanger genomes version 4.0. Of the 13191 proteins

in the *S. mansoni* proteome, 323 were predicted to be essential using this method. This low number of essential predictions was due to two factors, the relatively low number of essential genes observed in *C. elegans*, and the large evolutionary distance between *C. elegans* and *S. mansoni* that restricted orthomcl to find only 39% of the proteins in *S. mansoni* sharing any co-orthology relationships with the *C. elegans* proteome.

3.3.3 Precedence filter

The 323 putative essential *S. mansoni* genes were searched against the protein targets in ChEMBL (Gaulton *et al.*, 2011) using BLAST+, (E-value cutoff of 1×10^{-03} and target coverage of >50%). Only those hits to ChEMBL targets that had at least one potent compound (<10nM) were considered. This filter reduced the *S. mansoni* set to just 24 genes (Table 3.2) that were predicted to be essential and have the potential to be inhibited by small-molecule compounds.

3.3.4 Prioritized targets in the literature

In the *S. mansoni* genome study by Berriman *et al.* (2009), they defined two sets of potentially druggable targets. Both sets were found by homology searching ChEMBL and DrugStore (a database the targets of FDA approved drugs) for precedented targets, with the sets subdivided into those which were homologs of precedented human drug targets (26 proteins) and those which were homologs to targets with precedented drug-like chemical matter associated (94 proteins). The overlap of our results and those of Berriman *et al.* (2009) are shown in Table 3.2. The overlap was small as the homology criteria by Berriman *et al.* (2009) were

much stricter (>50% sequence identity and >80% target coverage) than my own, and they did not filter by any essentiality method.

3.3.5 GO term analysis of the prioritized targets

Gene ontology (GO) annotations for *S. mansoni* proteins were provided by Berriman *et al.* (2009) (downloaded from `ftp.sanger.ac.uk/pub/pathogens/`). GO terms were mapped to the generic GO-slim ontology (`http://geneontology.org/GO_slims/`) using map2slim from the go-perl library (`http://search.cpan.org/~cmungall/go-perl`). GO annotations were available for 8756 of the 13191 proteins of the *S. mansoni* proteome, which was used as a background population term set. GO annotations were available for 23 of the 24 proteins in the prioritized set. Over-representation of terms in the prioritized targets set was calculated using the Ontologizer software (Bauer *et al.*, 2008), with settings "Parent-Child-Union" as described in Grossmann *et al.* (2007). Only GO terms with the root biological process (BP) or molecular function (MF) were considered. Those GO terms significantly over-represented (p-value <0.1) are shown in Table 3.3.

3.3.6 Preliminary *in vitro* analysis

Laboratory analysis of the 24 *S. mansoni* genes is being undertaken by Alessandra Guidi at LSHTM. Preliminary results of the first two targets to be treated with RNAi are reported here. At the time of writing, two of these targets (Smp_026560.2 and Smp_096310) have been successfully knocked down with RNAi, achieving greater than 90% mRNA reduction in the cytoplasm. The resulting phenotypes are shown in Figures 3.3 and 3.4, the wild-type control phenotype is

Gene	Description	26set	94set
Smp_008260	serine/threonine kinase (CMGC group 3) (gsk 3-related)		yes
Smp_080730	serine/threonine kinase (CMGC group 3)		yes
Smp_096310	serine/threonine kinase (AGC group 5)		
Smp_141380	serine/threonine kinase (CK1 group 1)		
Smp_180400	serine/threonine kinase (CK1 group 2)		
Smp_009030	ribonucleoside-diphosphate reductase, alpha subunit, putative	yes	
Smp_026560.2	calmodulin, putative	yes	
Smp_027880	prefoldin subunit, putative		
Smp_034670	tubulin gamma chain, putative		
Smp_035580	protein phosphatase-1, putative		
Smp_040770	methionine-tRNA synthetase, putative		
Smp_041600	isoleucine-tRNA ligase		
Smp_055890	ribonucleoside-diphosphate reductase small chain, putative	yes	
Smp_073410	proteasome catalytic subunit 2 (T01 family)		
Smp_076230	proteasome subunit alpha 7 (T01 family)		
Smp_085740	abl-binding protein-related		
Smp_089700	integrin beta subunit, putative		
Smp_091770	protein farnesyltransferase alpha subunit, putative		
Smp_157090	subfamily C1A unassigned peptidase (C01 family)		yes
Smp_164840	proteasome catalytic subunit 3 (T01 family)		
Smp_165490	protein phosphatase-2a, putative		yes
Smp_170730	proteasome subunit alpha 1 (T01 family)		
Smp_173810	protein phosphatase pp2a regulatory subunit B, putative		yes
Smp_194160	leucyl-tRNA synthetase, putative		

Table 3.2: The 24 prioritized *S. mansoni* targets (predicted essential and with precedented active compound(s) available). Targets that were prioritized by Berriman *et al.* (2009) are highlighted, the 26set includes the targets homologous to human drug targets and the 94set includes the targets homologous to ChEMBL targets with drug-like chemical matter associated. Target annotation was taken from GeneDB (`http://www.genedb.org/`) and kinases were classified into subfamilies using the Kinomer v1.0 HMM library (Miranda-Saavedra & Barton, 2007).

shown in Figure 3.2. As it is not possible to reproduce the life-cycle of *S. mansoni* *in vitro*, it was not possible to deduce definitively that these genes were essential for disease progression or reproduction. However, in the opinion of the experts at LSHTM, these trematodes were significantly damaged, and potentially unviable.

Figure 3.2: Control phenotype of *S. mansoni*, no RNAi treatment. 10 days phenotype.

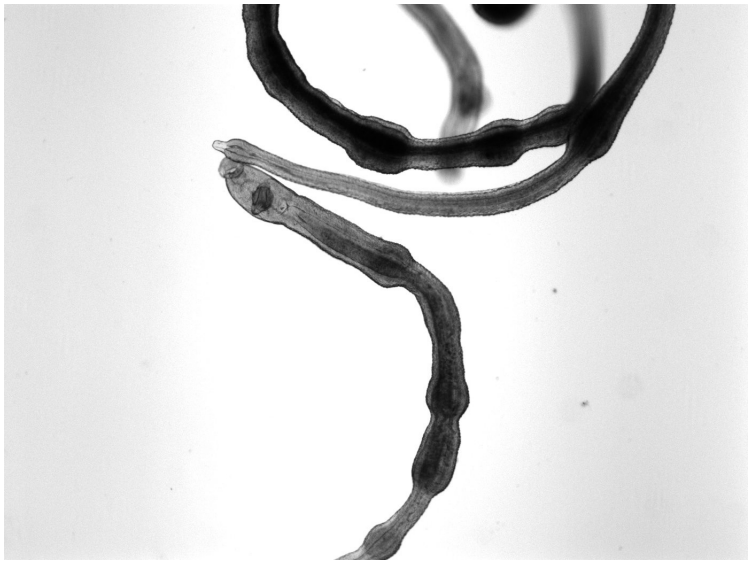


Figure 3.3: *S. mansoni* treated with RNAi designed against Smp_026560.2 (putative calmodulin). 10 days phenotype. Trematodes exhibit a severely segmented morphology and reduced motility.

GO id	GO name	*p*-value	GO root	Study Count	Population Count
GO:0006399	tRNA metabolic process	0.005	BP	3 (13.0%)	125 (1.4%)
GO:0016791	phosphatase activity	0.010	MF	3 (13.0%)	133 (1.5%)
GO:0016301	kinase activity	0.017	MF	5 (21.7%)	466 (5.3%)
GO:0009056	catabolic process	0.025	BP	6 (26.1%)	704 (8.0%)
GO:0006464	cellular protein modification process	0.026	BP	6 (26.1%)	713 (8.1%)
GO:0008233	peptidase activity	0.027	MF	4 (17.4%)	349 (4.0%)
GO:0016874	ligase activity	0.051	MF	3 (13.0%)	253 (2.9%)
GO:0006520	cellular amino acid metabolic process	0.069	BP	3 (13.0%)	162 (1.9%)
GO:0044281	small molecule metabolic process	0.081	BP	6 (26.1%)	929 (10.6%)
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	0.084	MF	1 (4.3%)	26 (0.3%)

Table 3.3: GO term over-representation analysis of the prioritized *S. mansoni* targets. The study and population counts show the number of targets each term is associated with in the prioritized set and the *S. mansoni* proteome respectively. The GO root shows the GO domain of each term, either Biological process (BP) or Molecular function (MF). The *p*-value was calculated by Ontologizer (Bauer *et al.*, 2008) using Fisher's Exact Test.

3.3.6.1 Discussion

There is currently no large-scale experimentally derived information on the essential genes in *S. mansoni*. In order to predict the essential genes using the methods described in Chapter 2, a reference species was required. There was limited choice of eukaryotic multi-cellular species with whole genome essentiality screens available, these included *Danio rerio* (zebrafish) (Amsterdam *et al.*, 2004), *Drosophila melanogaster* (fruit fly) (Boutros *et al.*, 2004), *Mus musculus* (house mouse) (Eppig *et al.*, 2012) and *C. elegans* (roundworm) (Kamath *et al.*, 2003). All of these species belonged to a different phylum to *S. mansoni*. *C. elegans* was used as it was deemed to be the most similar in life-cycle to *S. mansoni*,

Figure 3.4: *S. mansoni* treated with RNAi designed against Smp_096310 (serine/threonine kinase - AGC group 5), 22 days phenotype. Trematodes exhibit tegument damage and extremely reduced motility.

however it was evolutionarily very distant. This large distance was apparent in the orthology analysis, where only 39% of *S. mansoni* targets shared a detectable co-ortholog with *C. elegans*. The large distance most likely affected the functional types of targets which were predicted essential, as those targets that were maintained across such a large evolutionary distance were more likely to be involved in core biological processes. The GO term analysis showed that the biological processes overrepresented in the prioritized set were largely involved in these core functions, such as tRNA metabolism, catabolism and amino acid metabolism. The bias in target selection towards core metabolism could have implications on the need for pathogen-host selective drugs, as the core metabolism targets are also likely to be essential in the human host. Given the relatively small number of essential targets predicted in *S. mansoni* (323), it would be preferable to in-

fer essentiality from all species with any known essential genes, to increase the number of predictions.

Using ChEMBL it was intended to predict those targets that were likely to be inhibited by small molecule compounds. Of the 323 predicted essential targets only 24 showed significant sequence similarity to a ChEMBL target associated with potent compounds. This lack of targets may represent a true reflection of the druggability of the *S. mansoni* genome. However, as the vast majority (>85%) of ChEMBL targets are mammalian (largely human, mouse and rat), so it could also be that there is little experimental data available for non-mammalian targets. The GO term analysis showed that the molecular functions overrepresented in the prioritized set were enzyme functions such as phosphatases, kinases, peptidases and ligases. This was unsurprising as enzymes are more amenable to small molecule inhibition, due to their substrate binding sites often naturally accommodating small molecules.

It is possible to apply the essentiality models to multicellular organisms, however, at this time with so few experimental results and the lack of a clear essential phenotype, it is difficult to assign any confidence to the models. This simple procedure, of filtering a genome by two features, essentiality and a precedent for potent chemical matter, shows markedly how quickly potential target space can be diminished. Of the 11,809 genes of a human pathogen, only 24 were deemed suitable for a drug discovery program. Given that this process did not even consider host-pathogen selectivity issues, the 24 genes could be reduced even further. In the context of this collaboration, this was not an issue, however in a major drug discovery program, the chances of such a small number of target candidates progressing further would be small.

Chapter 4

Domain-based Inference for Druggability

4.1 Introduction

Druggability, the ability of a protein to bind "drug-like" small molecules with high affinity, is an important attribute to consider when prioritizing potential drug targets in a pathogen genome. However for druggability assessment of a pathogen proteome only a handful of proteins have published data on interactions with biologically active small molecules. For newly sequenced genomes of emerging pathogens there may be no known binding data or known compounds. To overcome the sparse availability of pharmacological data for proteins from pathogenic organisms, we can harness evolutionary information to infer chemogenomic druggability via homology. That is, if we know that a particular protein binds drug-like chemical matter then, by inference, proteins that are evolutionarily related are also likely to bind drug-like chemical matter. The relationship between evolu-

tionary distance and ligand binding potential is a complex one. The confidence in such inference is inversely proportional to the evolutionary distance between the pathogen protein and the homolog with known bioactivity. Importantly, such inference does not rely on homologous proteins necessarily retaining the capacity to bind a particular small molecule ligand, rather merely the capacity to bind any small molecule ligand - a much more conservative assertion.

4.1.1 ChEMBL homology for Druggability

The ChEMBL database (Gaulton *et al.*, 2011) (http://www.ebi.ac.uk/chembl/) comprises binding, functional and ADMET (absorption, distribution, metabolism, excretion and toxicity) information for drug-like bioactive compounds abstracted from more than 48,000 papers from the medicinal chemistry literature covering a period of more than 30 years. The data are manually abstracted from the primary literature and standardized. Features such as the molecular target, which are listed in the literature under numerous adopted names and synonyms are mapped to a non-redundant set of molecular targets, activity units are transformed into a standard preferred format (e.g. nM from μM for IC_{50}) and compounds are drawn in a machine-readable format (Gaulton *et al.*, 2011). The data is provided in a relational database format supporting multiple RDBMS platforms (such as Oracle, MySQL and PostgreSQL). To assess the likely druggability of proteins in a pathogen proteome one approach is to identify which pathogen proteins are homologous to proteins with established biologically active small molecule compounds (Aguero *et al.*, 2008; Berriman *et al.*, 2009). Such evolutionary relationships can be determined with standard sequence similarity

search algorithms such as BLAST+ (Camacho *et al.*, 2009). However, given the evolutionary distance, sequence similarity and orthology methods may be conservative in identifying druggable protein targets. Conversely proteins which contain multiple domains may be similar with drug targets in ChEMBL yet lack the vital domain which contains the drug binding site. A broader approach to identification of drug targets in a genome was proposed by Hopkins & Groom (2002) who suggested most drug binding sites can be mapped to proteins domains or specific configurations of domains to classify drug targets into druggable domain families, with conserved architectures (Overington *et al.*, 2006). The presence of a ligand binding site across domain families could then been assessed by structure-based sequence alignment within the family. However little work has been done in applying a domain-based approach to the mining of druggable targets in genomes since the original Hopkins & Groom (2002) publication. In this chapter a search method is described to identify a putative set of druggable domains for mining genomes. In the following chapter we describe how the binding sites in a druggable domain family can be analyzed across the entire family.

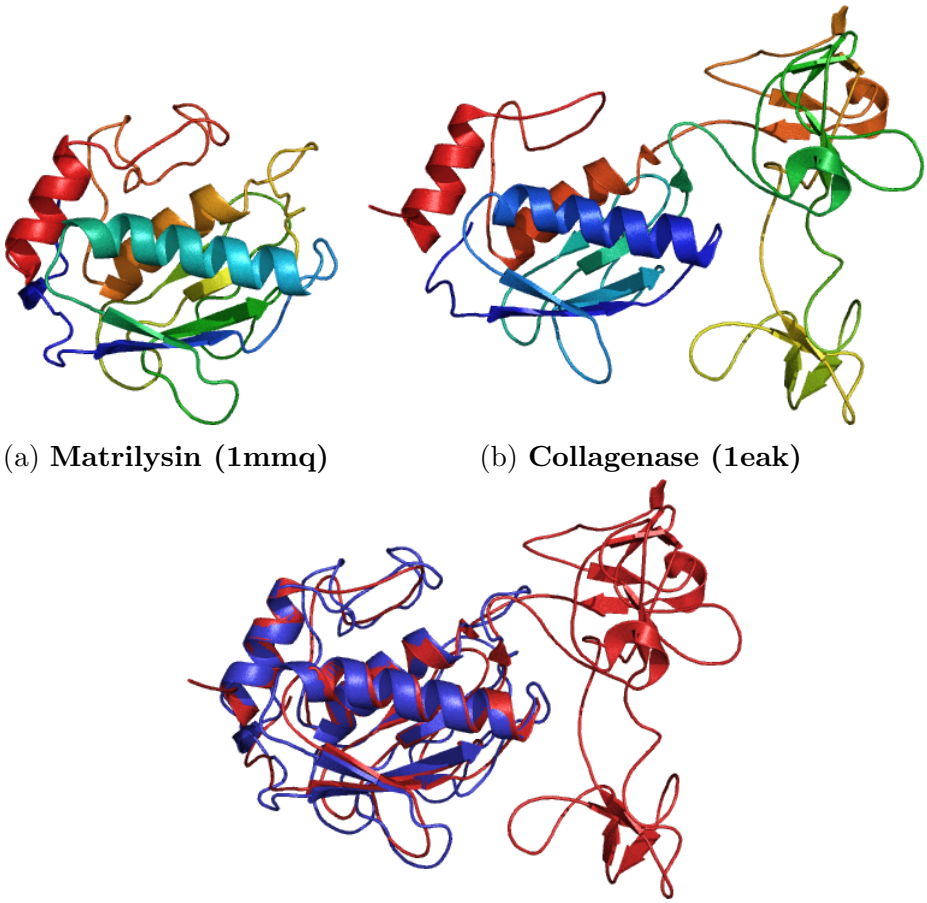
The chemogenomic druggability module of RAPID (rapid analysis of pharmacology for infectious diseases) harnesses the large-scale bioactivity data from ChEMBL to infer druggability by the quality and diversity of chemical matter associated with related targets. However, the description of bioactivity as a simple compound-target pair oversimplifies the biological reality. Drug targets often comprise multiple structural domains, multiple protein components and even multiple binding sites. Rigorous domain assignment is an important step in resolving these issues

4.1.2 What is a domain?

Domains are the building blocks of proteins, and their smallest evolutionary independent units. A domain is a collection of structural motifs which arrange to form a stable three-dimensional structure. If a domain is cleaved out of the parent protein, it would usually still fold, and often maintain function. Domain stability is usually achieved by nature of a hydrophobic core, however some smaller domains often use metal ions and disulfide bridges for stabilization. Evolution has made extensive re-use of domains to produce new proteins with new functions. Domains from different proteins that share the same structural organization and show detectable amino acid sequence similarity are said to belong to the same "family". Where this structural similarity is observed but sequence similarity is not present, these domains are said to belong to the same "fold". An often neglected feature of domains is that they may be "discontinuous", and be the product of multiple peptide regions from within a gene (Figure 4.1). Jones *et al.* (1998) found that between 25-30% of observed structures contain a discontinuous domain, and Dengler *et al.* (2001) observed that 19% of domain family representatives in the 3Dee database (Siddiqui *et al.*, 2001) were discontinuous.

4.1.3 Why we need domain annotation

An important feature of many domain families (and to a lesser extent folds), is that the family members often have varying levels of conserved function. For example, the majority of the protein kinase family function by phosphorylating specific residues of other proteins. Each member of the family may act on one or more specific substrate proteins, but essentially the mechanism of action is the



(a) **Matrilysin (1mmq)** (b) **Collagenase (1eak)**

(c) **Superimposed**

Figure 4.1: Example of continuous (a) and discontinuous (b) domain in the same family (Matrix metalloproteases, catalytic domain). Rainbow color scheme (N-terminus blue to C-terminus red) applied to (a) and (b). In (c) both domains superimposed over the common domain. The discontinuous domain (b) has three small domains (Fibronectin type II module) inserted towards the C-terminus. Figure generated using PyMOL (Delano, 2006)

same, i.e. transferring a phosphate group from an ATP molecule to a serine, threonine or tyrosine residue in the substrate protein. The consequence of conserved function(s) is often conservation of the physicochemical properties at important position in the domain (e.g. the catalytic site or co-factor binding sites). These functionally important sites are often the target for drug design, and as such the sharing of properties at these sites can directly effect the potential for selectivity of a drug against members of a domain family within a genome. Conversely, potential polypharmacology targets may be elucidated by exploiting the similarity at these sites.

4.1.4 Structure based domain annotation

For a large number of proteins, the three-dimensional (3D) structure has been determined and deposited in the Protein Data Bank (PDB)(Berman *et al.*, 2000). The two publicly available databases SCOP (Structural Classification Of Proteins) (Murzin *et al.*, 1995), and CATH (Class,Architecture,Topology,Homologous) (Orengo *et al.*, 1997), provide detailed annotation of the domain delineation within these 3D structures. Both databases offer a hierarchical classification system which include equivalents for fold and family levels, and both deal with discontinuous domains. The CATH process is largely automated and as such provides a fast turnaround from PDB submission to domain classification. The SCOP process is largely achieved by manual inspection, and as such is of very high quality. However this quality comes with a cost, and the release of data is sporadic. The current release of SCOP (version 1.75), only contains all high quality PDB submissions up to February 2009.

4.1.5 Sequence based domain annotation

There exists multiple publicly available resources for the analysis of conserved domains based solely on sequence homology. The CDD (Conserved Domain Database) is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignments (MSAs) for ancient domains and full-length proteins (Marchler-Bauer *et al.*, 2007). Sequences may be annotated with these models using Reverse Position-Specific BLAST (RPS-BLAST), which is more sensitive than BLAST as the targets are sequence profiles of the alignment models, rather than the individual sequences. The Pfam (Protein families) database (Finn *et al.*, 2010) is created by a similar process; high quality, manually curated families MSAs are collected into PfamA, and automatically generated sequence clusters are collected into PfamB. The annotations of sequences with Pfam utilizes a hidden Markov model (HMM), which is sensitive enough to detect extremely distant evolutionary relationships. Both these resources have a larger coverage of domain space than the structural annotations, as they do not require a 3D structure representative for an annotation. While these databases are comprehensive, they are not always accurate in their domain boundary predictions, and there are many examples of Pfam and CDD domain annotations that span multiple structural annotations. The sensitivity of the search methods also makes it difficult to distinguish between family annotations and more distant superfamily of fold relationships.

4.2 Methods

To annotate domains in our sequences it was decided that the high quality domain data in SCOP would be most preferable. However the lack of structural coverage would be supplemented with the high quality sequence annotations from PfamA. The rationale for preferential selection of SCOP domain annotations over those from Pfam is that SCOP has a fundamental conceptual framework in the form of a rigorous definition of what comprises a structural domain and an exclusive, self-consistent classification of these domains into a hierarchy. The lack of such an unambiguous classification in Pfam can result in overlaps in the ensuing annotations that can be difficult to resolve. These overlaps may be due to Pfam profiles describing more than one structural domain (or partial domain) or profiles describing protein families at different levels of similarity (e.g. family and super-family). Pfam has the clear advantage of not being restricted to proteins of known structure and can therefore provide much greater coverage. Hence the hybrid approach used here, use the superior structural domain annotations of SCOP where available but complement it with the greater Pfam coverage elsewhere.

4.2.1 SCOP search

The Astral database (Chandonia *et al.*, 2004) provides SCOP domain sequences extracted from the PDB with sequence redundancy removed. The query sequence to be annotated was searched against the Astral95 sequence set (`http://scop.berkeley.edu/astral/`) with BLAST+. All significant hits were ranked according to ascending E-value. As BLAST is a search tool, its heuristic algorithm did not guarantee the most optimal alignment between query and hit se-

quences. To improve the accuracy of the domain boundary prediction, the Smith-Waterman algorithm (Smith & Waterman, 1981) was applied. Smith-Waterman is a dynamic programming algorithm that performs local sequence alignment, and is guaranteed to find the optimal local alignment between a pair of sequences (with respect to the gap-scoring and residue substitution system being used). Each full length hit (representing a full domain) was aligned back to the target sequence using the EMBOSS (Rice *et al.*, 2000) implementation of Smith-Waterman. The resulting alignment was scanned for the presence of gap regions, as any significant gap regions (greater than 30 amino acid in length) could potentially be an inserted domain(s). The resulting aligned regions (minus significant gaps) were retained as a set of domain-fragment boundaries. If the combined length of these boundaries was greater than 60% of the target domain's full length, then the domain assignment was retained. As many protein sequences contain repeated domains, and as Smith-Waterman was guaranteed to find the optimal alignment, any assigned domains were masked out of the target sequence, and the process repeated until no more significant domain assignments were found.

4.2.2 Pfam search

The PfamA database was obtained from `ftp.sanger.ac.uk/pub/databases/Pfam/current_release/` and compiled using hmmpress from the HMMER package. The pfam_scan tool was installed locally (Finn *et al.*, 2010) to search PfamA with our query sequences.

4.2.3 Combining the annotations

The domain assignments were ordered firstly by SCOP over PFAM, then % target domain coverage, then by alignment score. Domain annotations are then assigned sequentially from this ordered list, where domain boundaries do not overlap with previously assigned domain boundaries by more than 25% of the domain length. The ordering method ensures that our preferred structural annotations take precedence. By prioritizing those annotations which cover more of the target domain, the domain boundaries are likely to be more accurate.

4.2.4 Small unannotated regions

Domain lengths in SCOP range from 21 residues (Retrovirus zinc finger-like domains, family) to 1504 residues (RNA-polymerase beta-prime, family). However the peak of the domain length distribution is round 100-110 residues and more than 85% of all domains are less that 200 residues. The loop regions of a protein (those amino acids that do not fold into secondary structures), occur both inter-domain ("linker") and intra-domain. The vast majority of these loop region are less than 10 amino acids long, however loops may be longer than 50 amino acids (Martin *et al.*, 1995). Long loops especially on the surface of proteins are often very mobile, and due to the nature of X-ray crystallographic techniques, may not be visible in the resulting structure. The outcome of these factors is that regions of a sequence that are unannotated could be large loops, linker regions or small domains, and it is not necessarily dependent on their length.

4.2.5 ChEMBL database

ChEMBL is a database of binding, functional and ADMET information of compounds extracted from published literature (Gaulton *et al.*, 2011). The version used in this work (version 01) contained over 400,000 distinct compounds assayed against over 5,500 (redundant by species) targets. Assay targets in ChEMBL can be a cell-line, organism, nucleic-acid or a protein. The only targets of interest for this analysis were the protein targets, of which there were 3,622. ChEMBL contains a broad range of assay endpoints but for this analysis the primary concern were those assays describing binding affinity or their surrogates. Only molecular target binding assays, whose endpoint was K_{i}, K_{d} or IC_{50} were selected. For druggability assessment, only compounds that bind potently were of concern, those having a binding affinity of greater than 10μm were excluded. However, as this carried the risk of excluding small but highly efficient binders, compounds whose binding affinity is greater than 10μm were also retained if they had high ligand efficiency (>0.3) (Hopkins *et al.*, 2004) as calculated by Equations 4.1 and 4.2. The free energy of ligand binding (Kuntz *et al.*, 1999) was defined at 300K using:

$$\Delta \mathrm{G} = -RT \ln K_{\mathrm{d}} \tag{4.1}$$

where R is the gas constant, T is the absolute temperature and K_{d} is the dissociation constant.

The ligand efficiency (Hopkins *et al.*, 2004) was calculated using:

$$\Delta \mathrm{g} = \Delta \mathrm{G} / \mathrm{N}_{\text{non-hydrogen}} \tag{4.2}$$

where N is the number of non-hydrogen atoms.

Each ChEMBL protein target was assigned as druggable if it was associated with at least one compound with the potency described above. The ChEMBL database was installed locally as an Oracle database for all the processing undertaken in this work. Note that the preparation and post-processing of the ChEMBL database was undertaken collaboratively, by myself and other members of the Andrew Hopkins group (University of Dundee).

4.3 Analysis

The ChEMBL protein targets were annotated with SCOP and PFAM domains, using the procedure described in section 4.2.1. Figure 4.2 shows the proportion of ChEMBL protein residues covered by domain annotation. Assuming a conservative linker length, of up to 20 residues, structural annotations cover 55% of ChEMBL protein residues. By adding sequence-based annotation, this figure increases to 72% coverage. If a less conservative linker length of 100 residues was allowed, then the protein residue annotation coverage increased to 80%. As SCOP is updated to reflect the rapidly growing PDB, the sequence coverage should improve as well.

Table 4.1 shows the distribution of domain complexity for the ChEMBL targets. 51% of the targets are single domain proteins, 25% have two domains and the remainder more than two. This distribution has important implications for attempts to assign ChEMBL compounds to structural domains, as nearly 49% of ChEMBL targets are multi-domain, and as such the domain associated with the activity of the targets compounds cannot be easily assigned.

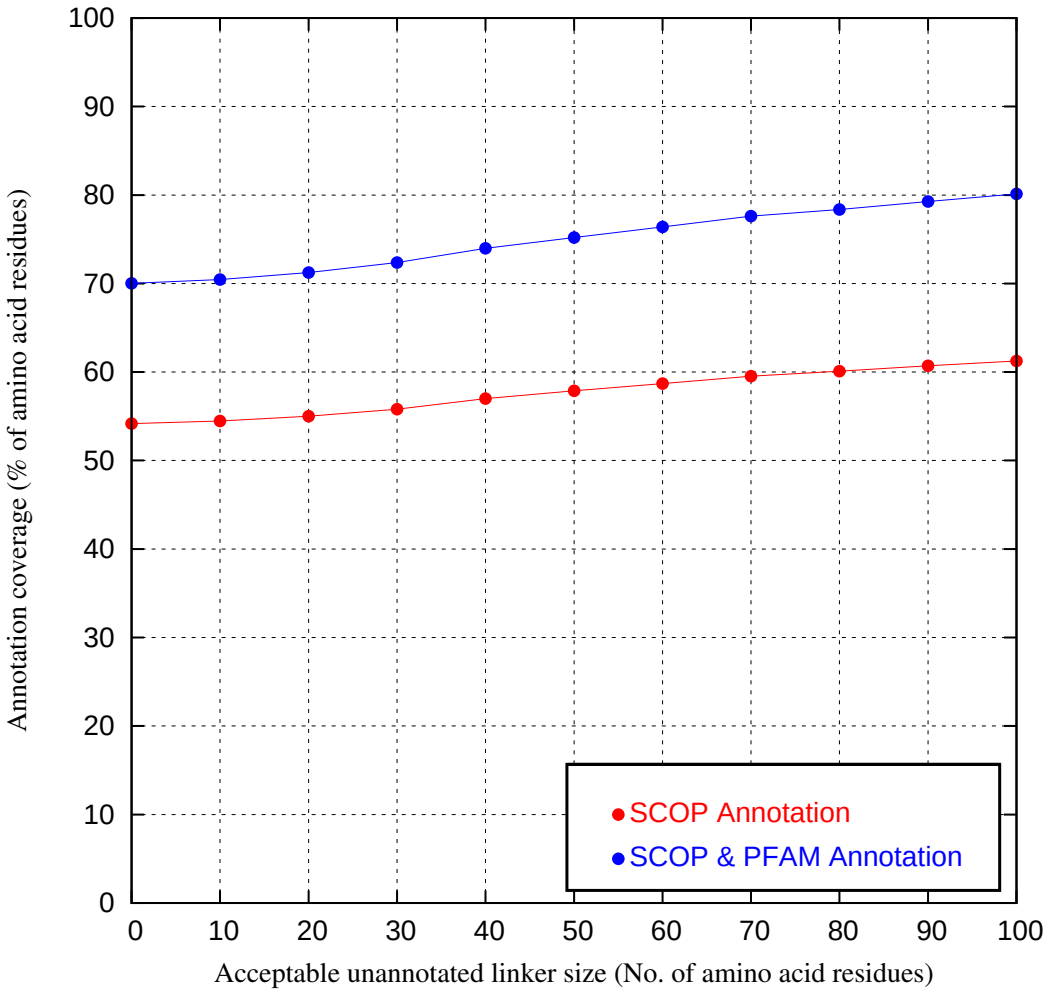


Figure 4.2: Domain annotation coverage of ChEMBL protein targets. The **annotation coverage** was calculated as the proportion of amino acid residues in the ChEMBL targets, annotated with a domain. Any consecutive residues without domain annotation were considered un-annotated, if their combined length was larger than the **acceptable unannotated linker** size. The structural annotation (SCOP) coverage is shown in red, and the increased coverage achieved by adding sequence-based (PFAM) annotation is shown in blue.

	ChEMBL protein targets		
Domain complexity	**Frequency**	**Coverage (%)**	**Cumulative coverage (%)**
1	1857	51.3	51.3
2	904	25.0	76.2
3	411	11.3	87.6
4	204	5.6	93.2
5	106	2.9	96.1
6	69	1.9	98.0
7	24	0.7	98.7
8	17	0.5	99.2
9	6	0.2	99.3
10+	24	0.7	100.0

Table 4.1: Distribution of domain complexity in ChEMBL protein targets. Over 50% of the targets are single domain proteins and less than 25% have more than two domains.

Table 4.2 shows the frequency distribution of the domain annotations in ChEMBL. The most frequent domains families are the Rhodopsin-like GPCRs, followed by the Protein Kinases and Ion Channels, all key drug target families. The related distribution in Table 4.3 shows the frequency distribution of "domain fingerprints" (DFP), which represent the canonically ordered combination of domains of a protein. In cases where the DFP consists of a single domain, ChEMBL compounds can be assigned to that domain with high confidence (at least in cases that lack significant unannotated regions). In many cases, where there are multiple domains in the DFP, the specific domain that is bound by the compound(s) is well characterized. For example, The Nuclear receptor ligand-binding domain (NRLBD) is almost always associated with a Nuclear receptor domain (NRD)(rank 7 in Table 4.3). The NRLBD is a well known druggable domain which is the target of tamoxifen (Shiau *et al.*, 1998), and other members of the NRLBD family are the targets for $\approx 13\%$ of all FDA approved drugs (Overington *et al.*, 2006). Therefore, the vast majority (if not all) of the compounds

screened against this DFP, will effect via the NRLBD domain. For the majority of the most common DFPs in ChEMBL, the ligand binding mechanisms are known and well studied. However, the number of activities per DFP in ChEMBL follow a power law, and for the many DFPs in the tail of the distribution, the ligand binding mechanism will be unknown.

Of the 525,801 ChEMBL activities used in this work 60% can be assigned using the top 22 DFPs, as shown in Table 4.4. Furthermore, 50% of the top activities are covered by the top 11 DFPs alone.

Figure 4.3 illustrates some of the more sophisticated visualizations that can be performed using these data. In this graph a node represents a domain family and an edge the co-occurrence of the connected domain families in at least one ChEMBL protein. A close up of the Giant Component is shown in Figure 4.4. The analysis illustrates that whereas the protein kinases co-occur with a broad range of different domain families, the Rhodopsin-like GPCRs and Ion Channels exhibit a more modest set of connections, and the Nuclear Hormone Receptor Ligand Binding Domain more modest still, and therefore more simpler to assign a compounds activity to a domain on a 1:1 basis.

4.3.0.1 Domain fingerprint over-representation

The most common DFPs in ChEMBL are mainly well characterized, eukaryotic, druggable targets. The proteins targets of ChEMBL are dominated by mammalian targets (>80%), and bacterial proteins represent < 10% of all the targets. In order to investigate the extent of bacterial-oriented information in ChEMBL, the domain annotation process was applied to the UniProtKB version of the *E. coli* K12 proteome. The inferred DFPs were then compared to the *E. coli* K12

R	Domain family	Td
1	Rhodopsin-like	376
2	Protein kinases, catalytic subunit	367
3	*Ion transport protein*	195
4	*Neurotransmitter-gated ion-channel transmembrane region*	92
5	*Neurotransmitter-gated ion-channel ligand binding domain*	84
6	I set domains	84
7	Fibronectin type III	82
8	Eukaryotic proteases	81
9	Nuclear receptor ligand-binding domain	73
10	Nuclear receptor	72
11	Adenylyl and guanylyl cyclase catalytic domain	67
12	SH2 domain	65
13	Cytochrome P450	59
14	Protein kinase cysteine-rich domain (cys2, phorbol-binding domain)	54
15	Neurotransmitter-gated ion-channel transmembrane pore	54
16	SH3-domain	48
17	PDEase	47
18	*Sushi domain (SCR repeat)*	47
19	*Fibronectin type III domain*	47
20	*7 transmembrane receptor (rhodopsin family)*	46
21	Phosphate binding protein-like	44
22	EGF-type module	43
23	Nicotinic receptor ligand binding domain-like	42
24	*Receptor family ligand binding region*	38
25	Hexokinase	38
26	L-arabinose binding protein-like	36
27	Higher-molecular-weight phosphotyrosine protein phosphatases	35
28	Histone deacetylase, HDAC	34
29	*RyR domain*	34
30	Carbonic anhydrase	34
31	PLC-like (P variant)	33
32	SNF-like	29
33	*RIH domain*	29
34	GAF domain	28
35	Complement control module/SCR domain	28
36	cAMP-binding domain	28
37	Voltage-gated potassium channels	28
38	Papain-like	27
39	*7 transmembrane sweet-taste receptor of 3 GCPR*	27
40	beta-Lactamase/D-ala carboxypeptidase	27
41	Tyrosine-dependent oxidoreductases	27
42	Nucleotide and nucleoside kinases	27
43	Kringle modules	26
44	Pleckstrin-homology domain (PH domain)	26
45	Glutathione S-transferase (GST), C-terminal domain	25

Table 4.2: The 45 most frequent domain families in ChEMBL protein targets. **Td** shows the total occurrences in the targets. PFAM and SCOP domain family annotations are shown in italics and roman respectively. Ranked (**R**) by **Td**.

R	Domain fingerprint (DFP)	Tt	Ct	Cc
1	Rhodopsin-like	366	10.1	10.1
2	Protein kinases, catalytic subunit	131	3.6	13.7
3	Cytochrome P450	59	1.6	15.4
4	Eukaryotic proteases	57	1.6	16.9
5	*Neurotransmitter-gated ion-channel ligand binding domain* *Neurotransmitter-gated ion-channel transmembrane region*	47	1.3	18.2
6	*7 transmembrane receptor (rhodopsin family)*	46	1.3	19.5
7	Nuclear receptor Nuclear receptor ligand-binding domain	44	1.2	20.7
8	Neurotransmitter-gated ion-channel transmembrane pore Nicotinic receptor ligand binding domain-like	37	1.0	21.7
9	Carbonic anhydrase	33	0.9	22.6
10	Phosphate binding protein-like *Receptor family ligand binding region*	26	0.7	23.4
11	Tyrosine-dependent oxidoreductases	25	0.7	24.0
12	SNF-like	25	0.7	24.7
13	Protein kinases, catalytic subunit SH2 domain SH3-domain	24	0.7	25.4
14	Histone deacetylase, HDAC	23	0.6	26.0
15	Glutathione S-transferase (GST), C-terminal domain Glutathione S-transferase (GST), N-terminal domain	23	0.6	26.7
16	PDEase	22	0.6	27.3
17	**No domain annotation**	22	0.6	27.9
18	Tubulin, C-terminal domain Tubulin, GTPase domain	21	0.6	28.5
19	Protein serine/threonine phosphatase	20	0.6	29.0
20	*Neurotransmitter-gated ion-channel ligand binding domain* *Neurotransmitter-gated ion-channel transmembrane region* **x 2**	20	0.6	29.6
21	Alcohol dehydrogenase-like, C-terminal domain Alcohol dehydrogenase-like, N-terminal domain	20	0.6	30.1
22	Nucleotide and nucleoside kinases	20	0.6	30.7
23	Vertebrate phospholipase A2	19	0.5	31.2
24	Adenylyl and guanylyl cyclase catalytic domain **x 2** *Domain of Unknown Function (DUF1053)*	19	0.5	31.7
25	Pepsin-like	19	0.5	32.2
26	L-arabinose binding protein-like *7 transmembrane sweet-taste receptor of 3 GCPR* *Nine Cysteines Domain of family 3 GPCR*	17	0.5	32.7
27	*Ion transport protein* **x 4** *Voltage gated calcium channel IQ domain*	16	0.4	33.2
28	Neurotransmitter-gated ion-channel transmembrane pore *Neurotransmitter-gated ion-channel ligand binding domain*	16	0.4	33.6
29	Dihydrofolate reductases	16	0.4	34.0
30	Fatty acid binding protein-like	15	0.4	34.5

Table 4.3: The top 30 most frequent domain fingerprints (DFPs) of ChEMBL protein targets. **Tt** shows the number of protein targets with the DFP. **Ct** shows the % coverage of targets by this DFP. **Cc** shows the cumulative coverage (%) of targets. PFAM and SCOP domain family annotations are shown in italics and roman respectively. Repeated domains are succeeded by the number of copies. The DFPs are ranked (**R**) by **Tt**.

R	Domain fingerprint (DFP)	Tt	Ta	Ca	Cc
1	Rhodopsin-like	366	135195	25.7	25.7
2	*Neurotransmitter-gated ion-channel ligand binding domain* *Neurotransmitter-gated ion-channel transmembrane region*	47	29798	5.7	31.4
3	Protein kinases, catalytic subunit	131	16989	3.2	34.6
4	*Neurotransmitter-gated ion-channel ligand binding domain* *Neurotransmitter-gated ion-channel transmembrane region* **x 2**	20	14381	2.7	37.3
5	Phosphate binding protein-like *Receptor family ligand binding region*	26	13163	2.5	39.8
6	Phosphate binding protein-like *N-methyl D-aspartate receptor 2B3 C-terminus* *Receptor family ligand binding region*	9	9699	1.9	41.7
7	*7 transmembrane receptor (rhodopsin family)*	46	9087	1.7	43.4
8	PDEase	22	8904	1.7	45.1
9	Eukaryotic proteases	57	8348	1.6	46.7
10	Neurotransmitter-gated ion-channel transmembrane pore Nicotinic receptor ligand binding domain-like	37	7854	1.5	48.2
11	Hemopexin-like domain MMP N-terminal domain Matrix metalloproteases, catalytic domain	10	7146	1.4	49.6
12	SNF-like	25	7081	1.3	50.9
13	Carbonic anhydrase	33	6733	1.3	52.2
14	EGF-type module Myeloperoxidase-like	12	6363	1.2	53.4
15	Cytochrome P450	59	5797	1.1	54.5
16	Eukaryotic proteases GLA-domain Kringle modules **x 2**	3	5080	1.0	55.5
17	EGF-type module **x 2** Eukaryotic proteases GLA-domain	8	4858	0.9	56.4
18	Histone deacetylase, HDAC	23	4519	0.8	57.2
19	Protein kinases, catalytic subunit SH2 domain SH3-domain	24	4353	0.9	58.1
20	SNF-like *Serotonin (5-HT) neurotransmitter transporter, N-terminus*	4	4349	0.8	58.9
21	Retroviral protease (retropepsin)	1	4287	0.8	59.7
22	Pepsin-like	19	4189	0.8	60.5

Table 4.4: The domain fingerprints (DFPs) that contribute 60% of all ChEMBL compound-protein activities. **Tt** shows the number of protein targets with the DFP. **Ta** shows the number of compounds with a reported activity against these targets. **Ca** shows the % coverage of activities by these targets. **Cc** shows the cumulative coverage (%) of activities. PFAM and SCOP domain family annotations are shown in italics and roman respectively. Repeated domains are succeeded by the number of copies. The DFPs are ranked (**R**) by **Ta**.

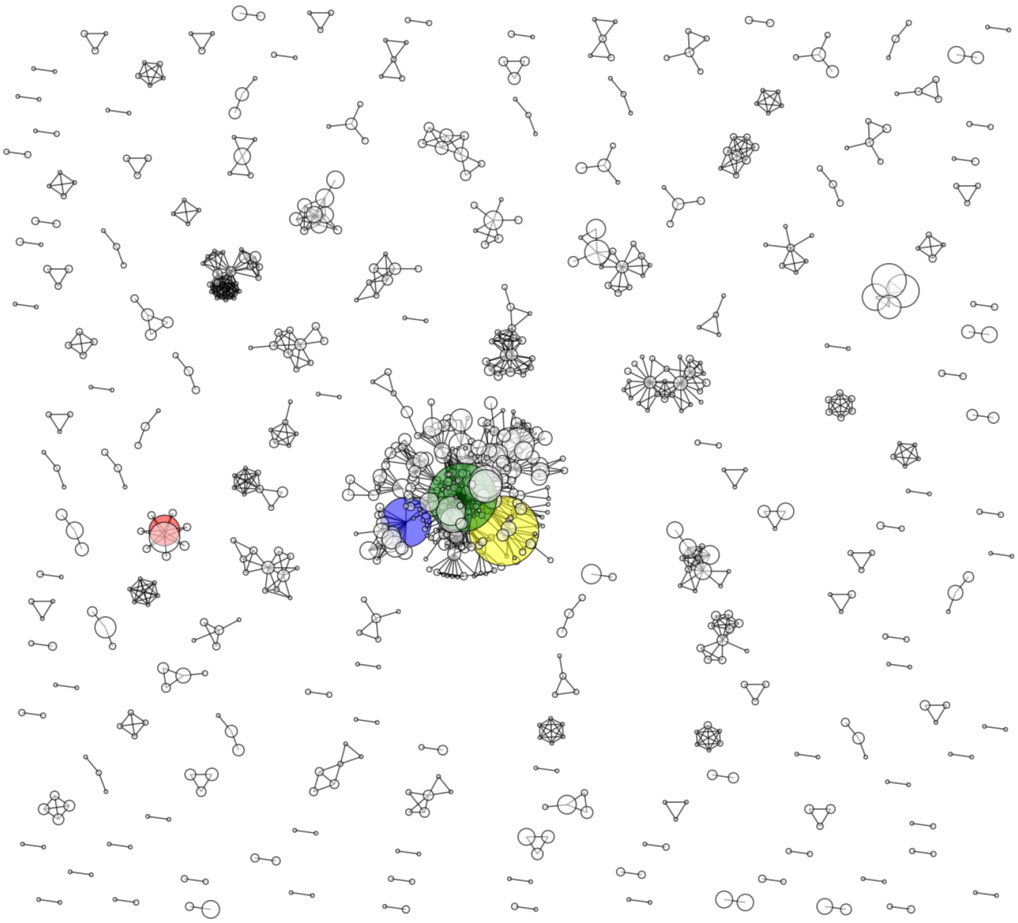


Figure 4.3: Domain co-occurrence graph for protein targets in ChEMBL. Nodes represent a domain family. Nodes share edges where they co-occur on a ChEMBL target. Node size is the number of occurrences. Unconnected nodes not shown. The Giant Component of the graph is expanded in Figure 4.3. Selected families colored as follows: *Protein kinases, catalytic subunit*: green, *Nuclear receptor ligand-binding domain*: red, *Rhodopsin like*: yellow and Ion trans: blue.

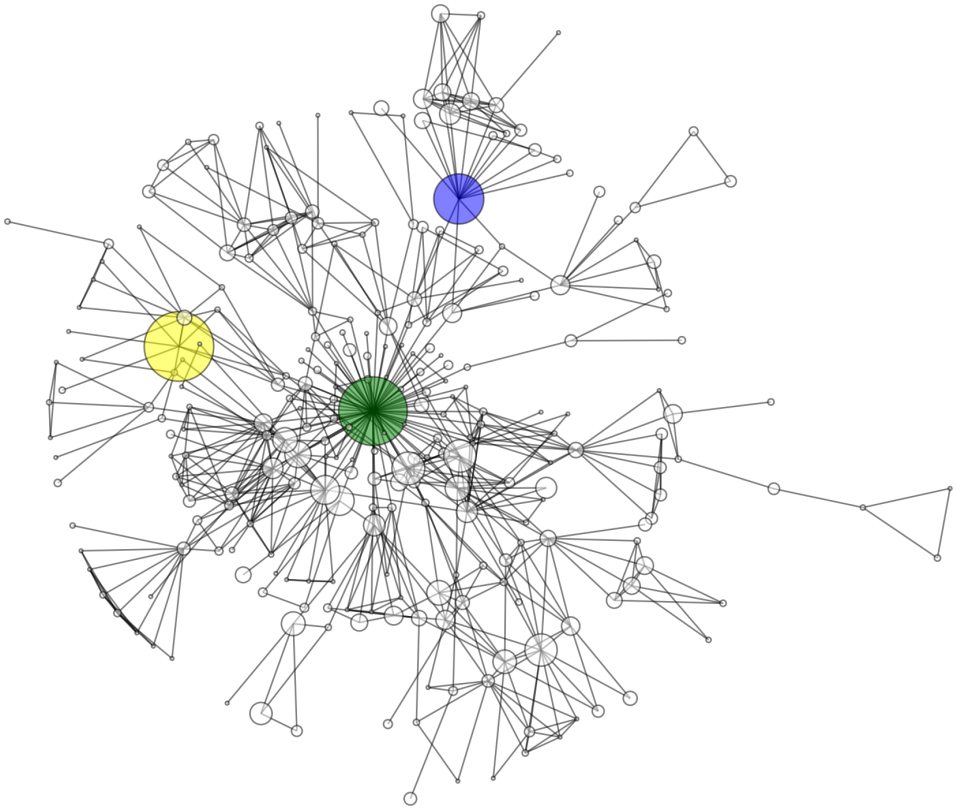


Figure 4.4: The Giant Component of network shown in Figure. 4.3. Domain co-occurrence of proteins in ChEMBL. Nodes represent a domain family. Nodes share edges where they co-occur on a ChEMBL target. Node size is the number of occurrences. Only largest (Giant) component shown. Selected families colored as follows: *Protein kinases, catalytic subunit*: green, *Rhodopsin like*: yellow and *Ion trans*: blue. Figure 4.3 and 4.4 were prepared using NetworkX (http://networkx.github.com/)

complement of ChEMBL, to find over-represented DFPs. *E. coli* was chosen as a representative of the bacteria as it has the most ChEMBL targets of all the prokaryotes. Significantly over-represented DFPs (p-value <= 0.01 using Fisher's exact test) are shown in Table 4.5. Of the 7 over-represented DFPs, the beta-Lactamase/D-ala carboxypeptidase domain was present in four of them. This domain is common in the family of penicillin binding proteins (the molecular targets of β-lactam antibiotics), which synthesize the essential peptidoglycan layer (Ghosh *et al.*, 2008). The Enolpyruvate transferases are the targets of the antibiotic fosfomycin (Kahan *et al.*, 1974). The targets *MurC* and *MurD* part of the peptidoglycan pathway and have been studied as potential drug targets (Zoeiby *et al.*, 2003), and DAHP synthetase is part of the shikimate pathway, which is essential for bacteria, but absent from mammals (Rizzi *et al.*, 2005).

***p*-value**	**Domain fingerprint (DFP)**	**ChOb**	**ChNt**	**PrOb**	**PrNt**
0.0001	PBP5 C-terminal domain-like beta-Lactamase/D-ala carboxypeptidase	3	69	3	4020
0.00017	Class I DAHP synthetase	3	69	4	4019
0.00017	beta-Lactamase/D-ala carboxypeptidase	3	69	4	4019
0.0018	Enolpyruvate transferase, EPT	2	70	2	4021
0.0018	MurCD N-terminal domain MurCDEF MurCDEF C-terminal domain	2	70	2	4021
0.0018	PBP transglycosylase domain-like beta-Lactamase/D-ala carboxypeptidase	2	70	2	4021
0.0018	Penicillin binding protein dimerisation domain beta-Lactamase/D-ala carboxypeptidase	2	70	2	4021

Table 4.5: The domain fingerprints (DFPs) significantly (p-value <= 0.01) over-represented in the ChEMBL *E. coli* targets, compared to the *E. coli* proteome. Where **ChOb** = observed in ChEMBL, **ChNt** = not-observed in ChEMBL, **PrOb** = observed in proteome and **PrNt** = not-observed in proteome.

4.4 Conclusions

The protein-ligand information available from ChEMBL provides an invaluable resource for inferring the likely druggability of related pathogen targets. However, description of bioactivity in terms of simple protein-ligand pairs oversimplifies the biological reality, drug targets often have more than one binding site, they frequently consist of multiple protein components and each component may consist of multiple structural domains. A pre-requisite for tackling the problem of assigning ChEMBL compounds to binding sites is the consistent domain annotation framework described here.

Currently, the ChEMBL database is greatly biased towards mammalian proteins, and a small set of domain families with a proven history of druggability. This bias could have a detrimental effect on inferring druggability onto prokaryotic pathogens. The lack of screening data for bacteria, is not a negative reflection on ChEMBL, which just reflects the publication content.

Reliable assignment of ChEMBL compounds to structural domains increases the accuracy of druggability inference approaches by reducing erroneous inferences through homology to non-binding domains. The work to improve the chemogenomic druggability inference, by putting it into the context of domain annotations is still ongoing, but the ChEMBL domain annotation procedures provided here represent a crucial first stage in achieving this.

Currently, given a pathogen proteome, those targets which share a common domain fingerprint with a ChEMBL protein target can be prioritized. Where multiple instances of potent compounds are associated with the ChEMBL target, the prioritization can be increased.

Chapter 5

Predicting Selectivity

5.1 Introduction

In a non-viral pathogen genome of interest, the intersect between targets that are both druggable and essential may be only 3% or less of the proteome, assuming these attributes are independent of each other. If broad-spectrum activity is required, this small percentage of possible drug targets may decrease even further. Anti-infective drugs usually need to be active against the pathogen proteins but selective over proteins in the host organism to reduce potential toxicity to the patient and adverse side effects. In searching for new anti-infective drug targets a common assumption has been that selectivity is determined by identifying proteins that are unique to the pathogen (Chan *et al.*, 2002) or are evolutionarily distant from any host protein (Kovalevskaya *et al.*, 2005).

To expand the limited number of potential drug targets that are essential, druggable and selective, the concept of selectivity can be refined. To be a drug target, essential pathogen genes do not have to be unique to the pathogen or ab-

sent from the host (Frearson *et al.*, 2007). Importantly, selectivity and therefore the required therapeutic index can be introduced by exploiting molecular differences in the binding site between homologous proteins present in both the host and pathogen proteomes (Zuccotto *et al.*, 2001). Analysis of protein sequences and structures may identify atomic differences in drug binding sites that could be exploited by medicinal chemistry to achieve selectivity and increase the therapeutic index. Toxicity, or other off-target undesired pharmacology that results from binding to a homologous human protein might be designed out if there are sufficient molecular differences between the homologous binding sites.

An example of how binding site differences between homologous proteins present in both human and pathogen genomes are exploited is shown by the antibacterial drug trimethoprim. Trimethoprim and other drugs in its class inhibit bacterial dihydrofolate reductase (DHFR) but not the related human DHFR enzyme. Trimethoprim binds bacterial DHFR many thousand times stronger than human DHFR (Hitchings, 1989). The mechanism of this selectivity is complex, but structural studies have shown that small differences in the binding sites of bacterial and human DHFR, result in the catalytic loop of the bacterial DHFR positioning much closer to Trimethoprim (Kovalevskaya *et al.*, 2007). In addition, the binding of co-factor (NADPH) in the bacterial DHFR results in hydrophobic interactions with Trimethoprim, which do not occur in human DHFR due to a greater distance between the ligands (Kovalevskaya *et al.*, 2005).

Selectivity can be manually assessed by comparing protein structures or models of homologous proteins. Using homology models has been shown to be an effective method of exploiting atomic differences between a pair of proteins to design selectivity into compounds (Hillisch *et al.*, 2004). This method is less effective

at the target selection stage, as many thousands of homologous protein pairs may exist between a pathogen and host. Methods have been developed to automate the generation of proteome-scale protein homology models for pathogen genomes (Aguero *et al.*, 2008; Ortí *et al.*, 2009; Pieper *et al.*, 2009). However, the generation of compound libraries, independent on the quality of the models, still requires individual comparison of models to assess selectivity. Moreover proteome-scale homology modeling requires intensive computational resources and quality control methods.

The GPRC SARfari and Kinase SARfari platforms https://www.ebi.ac.uk/chembl/, offer methods to compare ligand binding site distances between any number of family members, based on sequence metrics. These databases enable the user to visualize the proteins that share the "closest" binding site to a chosen target. However, there is yet no benchmark of the relationship between these binding site similarities and compound selectivity. A further limitation of this platform is that it is specific to just two protein families, and heavily focused on mammalian proteins. Detailed analysis of high-affinity kinase inhibitors by Sheinerman *et al.* (2005), showed that where kinases shared compound binding affinity, they had similar residues at specific positions important for binding. This method had the advantage that once the important binding positions are known for a compound, then the affinity of the compound could be inferred to any homologous protein from a pathogen or host. The drawback of this method was that it required a 3D structure of each compound complexed with a high affinity target, which are not available for the vast majority of screening compounds.

Currently, assessing the potential for selectivity between homologous human and parasite proteins is laborious and often experimentally intensive. Automated

metrics to predict selectivity are available, but their usefulness is limited due to the lack of any benchmark of their effectiveness.

5.1.1 Motivation

There is a need for high-throughput methods to assess the selectivity of orthologs and homologs within an organism and between organisms. Such methods would benefit a number of drug discovery strategies such as:

- **Drug re-profiling/repositioning**. Identifying binding sites in proteins which are similar to those of known drug targets. For example the human drug eflornithine was originally developed for facial hirsutism (Wolf *et al.*, 2007) but was also found to be an anti-trypanosomiasis drug (Pepin *et al.*, 1987). Eflornithine inhibits the enzyme ornithine decarboxylase in both humans and *Trypanosoma brucei*.

- **Polypharmacology**. Non-selective or "off-target" interactions can cause unwanted side effects. An example is the mode of action of the class of beta-blocker drugs, which reduce high-blood pressure by inhibiting β_1 receptors, but cause adverse effects due to their interactions with β_2 receptors (Bundkirchen *et al.*, 2003). Non-selective compounds have also been found to be beneficial in certain cases. Imatinib (gleevec), was designed to be a specific kinase inhibitor to combat chronic myelogenous leukemia, but studies showed it inhibited other human kinases and has since been approved for many other kinase dependent cancers (van Oosterom *et al.*, 2001). In pathogens, targeting multiple proteins simultaneously is both an effective way of expanding the "essentials" space, and potentially reducing the risk

of rapid resistance developing (Hopkins, 2008). It is important to note that polypharmacology can also exist in non-homologous targets, such as targets that share a common endogenous substrate or targets that share unexpected commonalities in their binding site (Hopkins *et al.*, 2011), and the work here does not attempt address these types of polypharmacology.

- **Selectivity**. Large-scale comparison of gene families shared between host and pathogen species to identify targets with binding sites with sufficient molecular differences to be exploited in drug design, to provide selective drugs.

5.2 Methods

In order to develop a generic methodology to predict selectivity between host and pathogen proteins it was decided to concentrate on a gene family approach based on comparison of drug binding sites within the domain. The previous chapter described a method to identify potentially druggable proteins from a proteome, based on their domain fingerprints. Here it is described how to compare domains within a domain family, whether the genes are from one or more species. In particular, to develop a method to compare the similarity of binding sites within a canonical gene family. To develop the method, the protein kinase family was selected. Protein kinases are one of the largest families of proteins in humans (Mayor *et al.*, 2004), and are present in many eukaryotic organisms. A solution to predicting kinase selectivity should therefore be applicable to most other gene families, where a drug/ligand binding site can be identified.

5.2.1 Protein Kinase Family

An accurate multiple sequence alignment of the domain family was required for multiple reasons as listed below:

- Transfer of ligand binding information from the PDB to the parent gene sequence.
- Comparing and condensing ligand binding sites across the domain family.
- Position specific profiles of the binding sites.
- Mapping/prediction of binding sites to all members of the family, as structural information is not always available.
- Enable sequence based calculations of binding site similarity.

The requirements of any such method would be:

- Family independent (protein kinases focus, but potentially any domain)
- Species independent (e.g. analyze a pathogen-ome)
- Extendable (e.g. add new structural data when available)

5.2.2 Seed Alignment

As described previously (Chapter 4), domains within the same family share detectable sequence homology. However these similarities are often localized to specific regions and residues and not uniformly spread over the domain. This often reflects in the difficulty of producing accurate sequence alignments, especially in these less conserved regions. The three-dimensional structure of a protein is

more conserved than sequence, and where available, these structures can be harnessed to improve the quality of sequence alignments. Structural information including solvent accessibility (buried in the globular protein or not), secondary structure assignment (α, π or 3_{10} helix, or β strand), uncommon hydrogen bonds (e.g. buried side-chain to main-chain amide), positive ϕ-torsion angles and disulphide bonds can assist the accurate alignment of residues within a domain family (Mizuguchi *et al.*, 1998a), where the sequences have diverged such that no obvious similarity remains. Where a protein family has a large coverage of structural information, a multiple structural alignment may be used as a seed, to create an alignment of the entire sequence family. A seed alignment is also valuable as a search tool, as profile-based methods are often more sensitive in retrieving more family members from a protein database.

Many methods exists to create structural alignments, such as STAMP (Russell & Barton, 1992), CORA (Orengo, 1999) and MISTRAL (Micheletti & Orland, 2009), as well as databases of pre-calculated structural alignments of protein families such as HOMSTRAD (Mizuguchi *et al.*, 1998b) and PALI (Balaji *et al.*, 2001). Despite the benefits of structural alignment, errors in alignment can still occur, especially in those structurally variable regions such as loops. In the example of the protein kinase family, the loop regions are variable in both size and due to mobility, in orientation. In these cases a combination of structural alignment, and expert manual adjustment is the preferred method. The Kinase SARfari database (`https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari`) provides a multiple alignment of protein kinase domains, created from a structural alignment with manual corrections. The alignment contained 959 unique sequences (Figure 5.1), representing the human kinome (including splice variants and poly-

morphisms) as well as orthologs from species with 3D structures available. The SARfari alignment was used as the seed alignment for the protein kinase family.

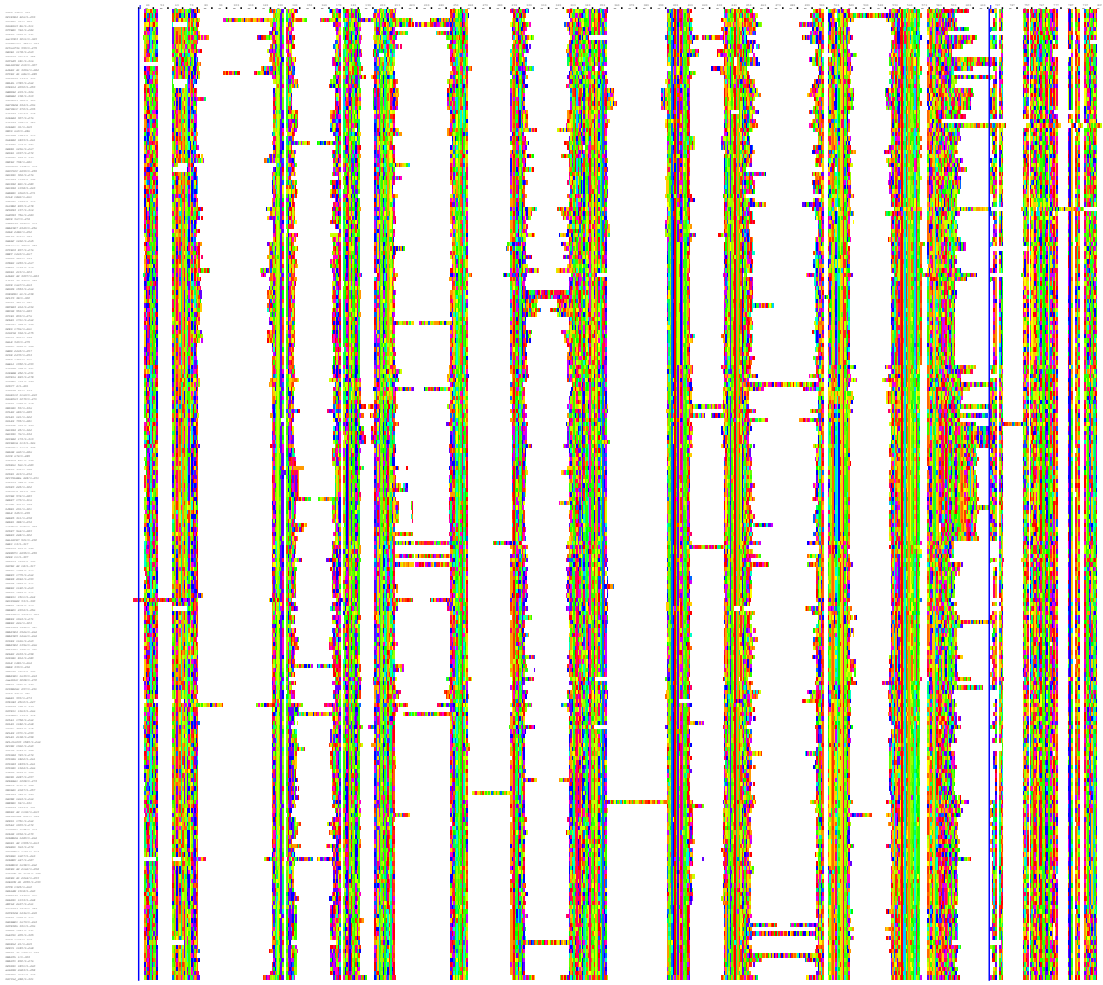


Figure 5.1: The multiple sequence alignment of the seed alignment (contains representative sequences of the Human Kinome, and all the parent genes of PDB kinases), created from the structural-based Kinase SARfari alignment. Sequence redundancy reduced to a maximum of 85% identical, resulting in 221 representatives of 959 kinase domains. Figure included only for illustrative purposes to highlight the size and complexity of this family. Visualization and sequence redundancy calculations performed using Jalview 2 (Waterhouse *et al.*, 2009)

5.2.3 Protein ligand binding information

"CREDO is a relational database storing all pairwise atomic interactions of inter- as well as intra-molecular contacts between small- and macromolecules found in experimentally-determined structures from the Protein Data Bank (PDB)" (http://marid.bioc.cam.ac.uk/credo). The ligand-protein atom interactions within crystal structures are pre-calculated by the CREDO process (Schreyer & Blundell, 2009). Multiple interaction types are calculated (see table 5.1), including hydrogen bonds, hydrophobic and aromatic interactions. In addition, CREDO also implements SIFTS (Structure Integration with Function, Taxonomy and Sequences) (Velankar *et al.*, 2005), to enable mapping of all polypeptide residues in the PDB onto their parent UniProt sequence, and therefore, a ligand to UniProt mapping. This is no simple matter as PDBs often contain chain breaks, modified residues, engineered inserts and mutations, all of which can reduce the accuracy of mapping. By using CREDO, a great deal of information on the observed ligand binding sites of protein kinases is already available, and the task is reduced to extracting and understanding this information.

5.2.4 Protein Kinases in the Human Genome - the "Kinome"

The complement of protein kinase domains within the human genome has been well studied by Manning *et al.* (2002). Whilst this is a valuable resource, the data had limitations, including domains that had a protein-kinase function but known not to be of the same structural family, the versions of the human genome studied, Ensembl IPI.1 (Lander *et al.*, 2001) and Celera 25h (Venter *et al.*, 2001)

CREDO name	Interaction type	Note
is_covalent	covalent bond	
is_vdw_clash	*van der Waals* clash	
is_vdw	*van der Waals*	
is_proximal	ligand-protein atom pair close (<6Å), and no other contacts	this contact type ignored in this work
is_hbond	hydrogen bond	
is_putative_hbond	putative hydrogen bond	
is_weak_hbond	weak hydrogen bond	
is_xbond	halogen bond	
is_ionic	ionic interaction	
is_metal_complex	metal complex	
	cation-π interactions	
is_pi_donor	donor	
is_pi_cation	cation	
is_pi_carbon	carbon	
	aromatic interactions	see Chakrabarti & Bhattacharyya (2007) for detailed explanation of aromatic interaction types
is_aromatic_ff	face-to-face	
is_aromatic_of	offset face-to-face	
is_aromatic_ee	edge-to-edge	
is_aromatic_ft	non-parallel face-to-face	
is_aromatic_ot	offset non-parallel face-to-face	
is_aromatic_et	offset non-parallel edge-to-face	
is_aromatic_fe	face-to-edge	
is_aromatic_oe	offset edge-to-face	
is_aromatic_ef	edge-to-face	
is_hydrophobic	hydrophobic interaction	
is_carbonyl	carbonyl interaction	

Table 5.1: The protein-ligand atom interaction types defined in the CREDO database (Schreyer & Blundell, 2009). The CREDO names shown here refer to the column names in the contacts table of the CREDO sql database (see Appendix A). All interaction types with the exception of is_proximal were used to define protein-ligand contacts.

were old (*c.* 2001), and there is potential for multiple gene sequence revisions and nomenclature changes. An alternative (Martin *et al.*, 2009) spans the kinomes of many species. An extra consideration is that most protein families have not been studied as rigorously as the protein kinases, and a more general method for retrieving the family complement from a genome is required. One method was described in Chapter 4. However as a high quality seed alignment was available for this family (see 5.2.2), the HMMER package (Eddy, 2011) was used to create a Hidden Markov Model (HMM) of the seed alignment. The HMM was then used to search against the UniProtKB (Magrane & Consortium, 2011) version of the human proteome to find the human protein kinase complement, the "Kinome". The HMMER (v3.0) suit of programs were installed locally. The seed alignment HMM was created using the hmmbuild program with default parameters, the human proteome HMM database was created by converting the FASTA file with hmmpress, and the HMM search was performed using hmmscan with a best domain hit E-value cutoff of 1×10^{-03} (-domE option), and the heuristic filters off (-max option). The kinase domain regions were extracted from the hmmscan results and used to created a kinase domain FASTA format sequence database representing the human Kinome.

5.2.5 Kinase Structures

To enable the use of CREDO ligand binding information, the complement of protein-kinases within the PDB was required. While databases such as SCOP accurately classify domains, they are not up-to-date. To collate the current kinase structures from within the PDB, the same procedure as applied to find the human

kinome (see 5.2.4), was applied to PDB protein sequences. As multiple PDB structures of the same kinase were present in the PDB, each kinase structure was mapped to its parent UniProt sequence using the residuemap table of the CREDO database. Representing the PDBs by their UniProt sequences, not only reduced redundancy, but also removed potential alignment errors caused by chain breaks (gaps in the protein sequence cause by unresolved residues), and enabled the removal of artificial fusion proteins. Many of the PDB kinase structures were not of human proteins (e.g. pdb 3dk3, structure of mouse *abl1* kinase), in these cases the parent UniProt protein sequence was retrieved and the domain region added to a FASTA format sequence database representing the kinase structures.

5.2.6 Family Alignment

The full multiple sequence alignment of the human kinome domains and the kinase structure domains was created using the seed HMM as the input for the hmmalign program of the HMMER package. The resulting family alignment was added to the matrix database (Figure 5.2). The boundaries of kinase domains within the full length UniProt sequences were recorded in the database, such that any position within the sequence alignment, could be mapped to the correct position within the parent UniProt sequence, and by querying the CREDO database, mapped to the correct residue in any PDB structure of that kinase. Conversely, any potential ligand binding residues in PDB structures could be mapped through the database to the equivalent residues in any human kinase.

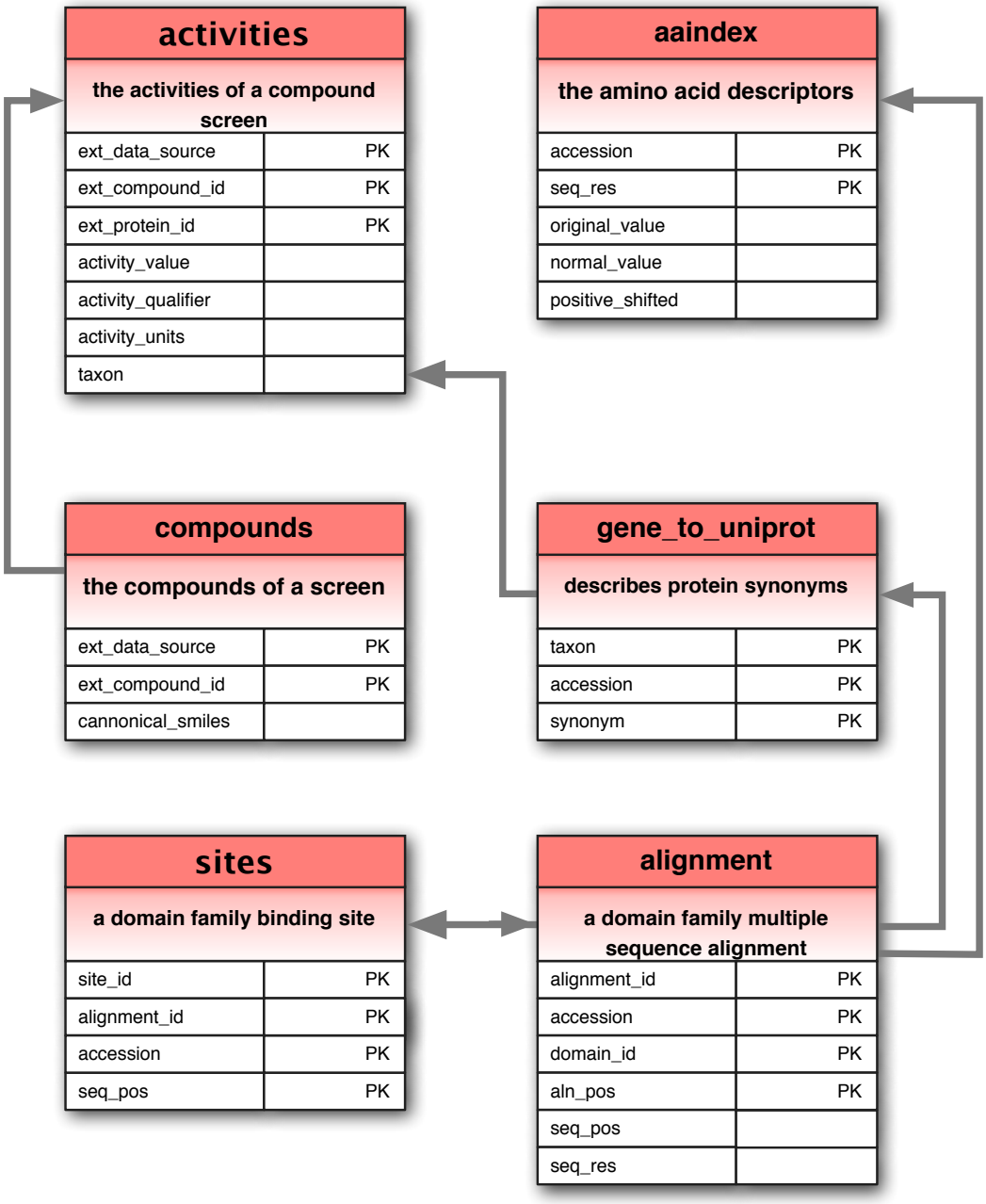


Figure 5.2: The matrix database schema for multiple sequence alignment, binding sites, activities and amino acid properties. The composite primary keys are represented by PK . Arrows indicate the direction of foreign key inheritance. For more details on column names and example queries see Appendix A.5. Figure generated using OmniGraffle (Case, 2013).

5.2.7 Amino acid properties

The AAindex database (Kawashima & Kanehisa, 2000) is a publicly available resource that collates published information on the observed and theoretical properties of amino acids. The current version of the database (v9.1) was obtained from ftp://ftp.genome.jp/pub/db/community/AAindex/. The database is divided into three sections, of which only AAindex1 was of interest. AAindex1 contains indices of numerical values for the 20 standard amino acids. Each index attempts to describe numerically, a specific physicochemical or biochemical property for each amino acid. Many of these indices relate to properties that could intuitively be important in ligand binding such as the *Hydrophobicity index* (ARGP820101) and the *Residue volume* index (GOLD730102), whereas others such as the *Normalized frequency of beta-sheet, with weights* (LEVM780102) pertain to features more commonly associated with protein structure. The current version contained 544 distinct indices, of which 529 contain a value for all 20 amino acids. The scale of the numerical values assigned to residues within an index is determined by the nature of the property measured, thus some indices contain very large values, while some contain small or negative values. To remove bias from the indices, each range was transformed to fit a defined range *Rmin* to *Rmax*:

$$X_{i,Rmin..Rmax} = \frac{(Rmax - Rmin)(X_i - X_{Min})}{X_{max} - X_{min}} + Rmin \tag{5.1}$$

where $X_{i,Rmin..Rmax}$ is the scaled value, X_i is the original value, X_{Max} is the maximum value within the index and X_{Min} the minimum value within the index. Values were calculated to both ranges [-1 to +1], and [1 to 10]. The complete indexes and normalized values were included in the matrix database (Figure 5.2).

5.2.7.1 Sheinerman Descriptors

Sheinerman *et al.* (2005) published a study of a handful of high-affinity protein kinase inhibitors and their kinase inhibition profile. They found that a small number of non-conservative amino acid substitutions at specific positions in the binding site could drastically reduce affinity. They produced rules to describe what constitutes a non-conservative amino acid substitution. These rules could be implemented using a small number of amino acid properties. By using equivalent descriptors from the AAindex database, a set of descriptors that described these important properties was produced (see Table 5.2), this set is referred to as a "descriptor model".

Sheinerman rule	**AAindex accession**	**AAindex description**
charge substitution	FAUJ880111 FAUJ880112	Positive charge Negative charge
polarity substitution	GRAR740102	Polarity
Van der Waals volume difference of $20\mathring{A}^3$	GRAR740103	Volume

Table 5.2: The AAindex descriptors used to define the *Sheinerman* descriptor model. The descriptor model based on rules of amino acid conservation that effect ligand binding.(Sheinerman *et al.*, 2005)

5.2.7.2 Westen Descriptors

van Westen *et al.* (2011) published a study on the effect of Reverse transcriptase mutations on inhibitor binding. In the study they selected indices from the AAindex that were descriptive of ligand binding properties or described structural constraints. A descriptor model was created using the 58 indices suggested (Table 5.3).

Table 5.3: The AAindex descriptors used to define the *Westen* descriptor model (van Westen *et al.*, 2011).

AAindex accession	AAindex description
ARGP820103	Membrane-buried preference parameters
BAEK050101	Linker index
BHAR880101	Average flexibility indices
CASG920101	Hydrophobicity scale from native protein structures
CHAM810101	Steric parameter
CHAM820101	Polarizability parameter
CHAM830101	The Chou-Fasman parameter of the coil conformation
CHAM830107	A parameter of charge transfer capability
CHAM830108	A parameter of charge transfer donor capability
CHOP780201	Normalized frequency of alpha-helix
CHOP780202	Normalized frequency of beta-sheet
CHOP780203	Normalized frequency of beta-turn
CIDH920105	Normalized average hydrophobicity scales
COSI940101	Electron-ion interaction potential values
FASG760101	Molecular weight
FAUJ880102	Smoothed upsilon steric parameter
FAUJ880103	Normalized van der Waals volume
FAUJ880104	STERIMOL length of the side chain
FAUJ880105	STERIMOL minimum width of the side chain
FAUJ880106	STERIMOL maximum width of the side chain
FAUJ880109	Number of hydrogen bond donors
FAUJ880110	Number of full nonbonding orbitals
FAUJ880111	Positive charge
FAUJ880112	Negative charge
FAUJ880113	pK-a(RCOOH)
GRAR740102	Polarity
JANJ780102	Percentage of buried residues
JANJ780103	Percentage of exposed residues
JOND920102	Relative mutability
JUNJ780101	Sequence frequency
KLEP840101	Net charge
KOEP990101	Alpha-helix propensity derived from designed sequences
KOEP990102	Beta-sheet propensity derived from designed sequences
KRIW790101	Side chain interaction parameter
KYTJ820101	Hydropathy index

Continued on next page

Table 5.3 – *Continued from previous page*

AAindex accession	AAindex description
LEVM760102	Distance between C-alpha and centroid of side chain
LEVM760103	Side chain angle theta(AAR)
LEVM760104	Side chain torsion angle phi(AAAR)
LEVM760105	Radius of gyration of side chain
LEVM760106	van der Waals parameter R0
LEVM760107	van der Waals parameter epsilon
MITS020101	Amphiphilicity index
MONM990201	Averaged turn propensities in a transmembrane helix
NISK800101	8 A contact number
NISK860101	14 A contact number
PONP800101	Surrounding hydrophobicity in folded form
PONP930101	Hydrophobicity scales
RACS770103	Side chain orientational preference
RADA880108	Mean polarity
ROSG850101	Mean area buried on transfer
ROSG850102	Mean fractional area loss
ROSM880102	Side chain hydropathy, corrected for solvation
TAKK010101	Side-chain contribution to protein stability
VINM940101	Normalized flexibility parameters (B-values), average
WARP780101	Average interactions per side chain atom
WOLR810101	Hydration potential
ZHOH040102	The relative stability scale extracted from mutation experiments
ZHOH040103	Buriability

5.2.8 Ligand Binding Sites

5.2.8.1 Installing CREDO

The current incarnation of CREDO is available as a queryable database using web services. At the time of this work, the database was only available as a mysql database. To enable cross querying of our internal database and CREDO, it was important to have an oracle version available. As most databases (CREDO included) use non-ANSI features of their platform, converting from one database

engine to another was a significant task. The schema of CREDO was available in mysql-specific SQL, in order to create the schema in oracle, this SQL was hand translated to the oracle equivalent (see Appendix A.5.2 for usage example and database schema). The CREDO data was added to the oracle database using sqlloader.

5.2.8.2 Extracting Ligand Binding Sites

Every kinase PDB-ligand pair in CREDO was queried and the residues with ligand interactions stored as a profile (protein-ligand interaction profile or PLIP), where the profile positions correspond to the residue positions in the family alignment. A metric called "contacts" was devised to weight those residues that may be more important for ligand binding. The metric was defined as the number of unique ligand-atoms, that shared at least one interaction (described in Table 5.1) with any of the residue's atoms. Figure 5.3 shows an example of the contacts metric for PDB entry 1k3a. The contacts number was used as the value in each of the PLIP positions (see Table 5.4). A PLIP was therefore a 2D representation of the ligand's binding site. As the PLIPs were based on alignment position, they enabled direct comparison or clustering of a ligand binding sites to all other ligand binding sites, from any of the other kinases (see Table 5.5). The PDB structures contained many ligands which were not of interest, such as buffers, very small ligands and unresolved ligand fragments. To reduce the impact of these ligands on the analysis, any ligand which had fewer than 8 heavy atoms or had fewer than four non-zero positions in its PLIP were removed.

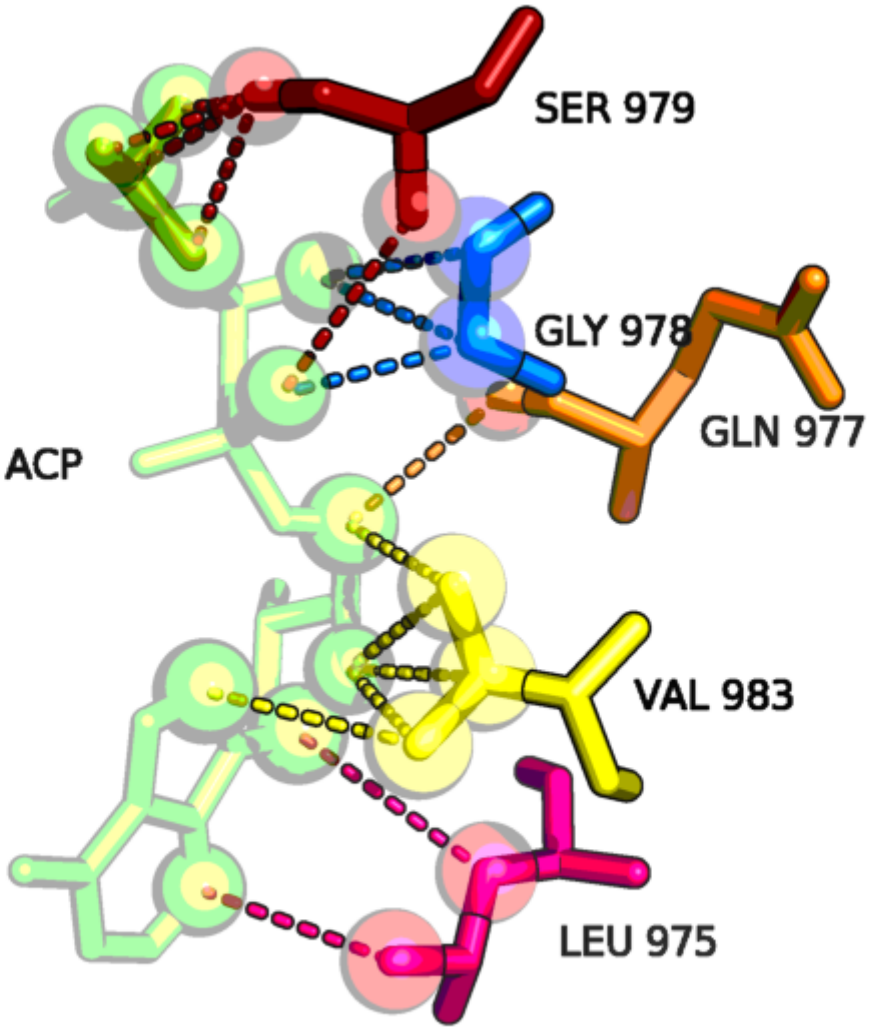


Figure 5.3: The contacts metric used to create PLIPs. An example using PDB entry 1k3a. Ligand ACP (β,γ-Methylene ATP) is shown in green. Spheres show atoms involved in interactions. Dotted lines connect residue-ligand atoms with an interaction. The residues shown here are the subset of residues with contacts in Table 5.4. An example of the contacts metric is VAL 983 (shown in yellow), which has 5 ligand-residue interactions with **3** distinct ligand atoms. Figure generated using PyMOL (Delano, 2006)

(MSA) Alignment position	UniProt residue number	Residue type	PDB residue number	Contacts
185	1005	**LEU**	**975**	**2**
186	1006	GLY	976	0
187	1007	**GLN**	**977**	**1**
188	1008	**GLY**	**978**	**2**
189	-	-	-	0
190	-	-	-	0
191	1009	**SER**	**979**	**5**
192	1010	PHE	980	0
193	1011	GLY	981	0
194	-	-	-	0
195	-	-	-	0
196	-	-	-	0
197	1012	MET	983	0
198	-	-	-	0
199	1013	**VAL**	**983**	**3**
200	1014	TYR	-	0

Table 5.4: PLIP (protein-ligand interaction profile) of Insulin-like growth factor 1 receptor (UniProt entry P08069) with ligand ACP (β,γ-Methylene ATP), constructed from PDB entry 1k3a. The alignment position refers to the Kinome multiple sequence alignment (see 5.2.6). The contacts count refers to the number of ligand-atoms involved in residue-ligand interactions (see 5.2.8.2), as shown in Figure 5.3. The PLIP shown here is a small section of the whole profile (positions 185-200 of 1769 positions).

(MSA) Alignment position	P51955 *NEK2* ADP (2W5A)	P08069 *IGF1R* ACP (1K3A)	P24941 *CDK2* ATP (1QMZ)	P24941 *CDK2* ATP (2CJM)	P24941 *CDK2* ATP (1FIN)	P24941 *CDK2* ATP (1JST)
185	1	2	5	0	6	2
186	0	0	0	3	2	6
187	0	1	1	1	5	3
188	3	2	1	1	4	6
189	0	0	0	4	0	0
190	0	0	0	0	0	0
191	1	5	1	0	3	2
192	0	0	0	2	0	3
193	0	0	0	0	0	0
194	0	0	0	0	0	0
195	0	0	0	0	0	0
196	0	0	0	0	0	0
197	0	0	0	0	0	0
198	0	0	1	0	0	0
199	5	3	3	3	2	2
200	0	0	0	0	0	0

Table 5.5: Multiple PLIPs (protein-ligand interaction profiles) of three kinases bound to three distinct ligands. Human *CDK2* bound to ATP (adenosine triphosphate) in four pdb entries, human *IGF1R* bound to ACP (β,γ-Methylene ATP) and human *NEK2* bound to ADP (adenosine diphosphate). The numbers in blue show the number of ligand-atoms involved in residue-ligand interactions (see 5.2.8.2) at that MSA position. The MSA position refers to the Kinome multiple sequence alignment (see 5.2.6). The PLIPs shown here are small sections of the whole profile (positions 185-200 of 1769 positions). This example shows some of the variations observed in ligand interactions. Where the same ligand (ATP) has been co-crystallized with the same protein (*CDK2*), the observed ligand interactions differ substantially.

(MSA) Alignment position	P24941 *CDK2* ATP (1QMZ)	P24941 *CDK2* ATP (2CJM)	P24941 *CDK2* ATP (1FIN)	P24941 *CDK2* ATP (1JST)	P24941 *CDK2* ATP Merged
185	5	0	6	2	3.25
186	0	3	2	6	2.75
187	1	1	5	3	2.50
188	1	1	4	6	3.00
189	0	4	0	0	1.00
190	0	0	0	0	0.00
191	1	0	3	2	1.50
192	0	2	0	3	1.25
193	0	0	0	0	0.00
194	0	0	0	0	0.00
195	0	0	0	0	0.00
196	0	0	0	0	0.00
197	0	0	0	0	0.00
198	1	0	0	0	0.25
199	3	3	2	2	2.50
200	0	0	0	0	0.00

Table 5.6: Merging PLIPs (protein-ligand interaction profiles). Four PLIPs describing the residue-ligand interactions of Human *CDK2* bound to ATP in four pdb entries, are merged into a single PLIP. Values in blue show the number of ligand-atoms involved in residue-ligand interactions (see 5.2.8.2) at that MSA position (or in the case of the merged PLIP value shows the averaged interactions at that position). The MSA position refers to the Kinome multiple sequence alignment (see 5.2.6). The PLIPs shown here are small sections of the whole profile (positions 185-200 of 1769 positions).

5.2.8.3 Clustering Ligand Binding Sites

The PDB contains multiple versions of the same kinase domains, often bound to the same ligand or very similar ligands. This resulted in many fully redundant ligand-kinase profiles, as well a large number of highly similar profiles. It was useful to collapse these similar profiles, which usually describe the same ligand-kinase binding site, with small discrepancies resulting from experimental errors and crystallographic resolution. Where profiles shared the same ligand, UniProt and overlapping binding positions, the profiles were merged, retaining the average number of contacts at each position (see Table 5.6). Further reduction in the number of ligand-kinase profiles was achieved with hierarchical clustering. Hierarchical clustering was performed using the Cluster 3.0 software (de Hoon *et al.*, 2004), with the Uncentered correlation metric and Average linkage setting. Manual inspection of the hierarchical clustering results was performed using TreeView (Saldanha, 2004). Appendix Figure A.5 shows the clustering tree and heatmap of the 739 merged PLIPs. Manually inspecting the clustering, it was possible to rediscover some of the known ligand binding modes observed in kinases (Figures 5.4 and 5.5). These ligand binding site definitions enabled mapping of any binding mode onto any kinase within the database.

5.2.8.4 Defining Ligand Binding Sites

As is shown in the hierarchical clustering (Appendix Figure A.5), where multiple structures are solved with the same, or highly similar ligands bound, the residue contact profile can vary significantly. These variations can be attributed to many factors, such as protein structure flexibility, ligand flexibility, conforma-

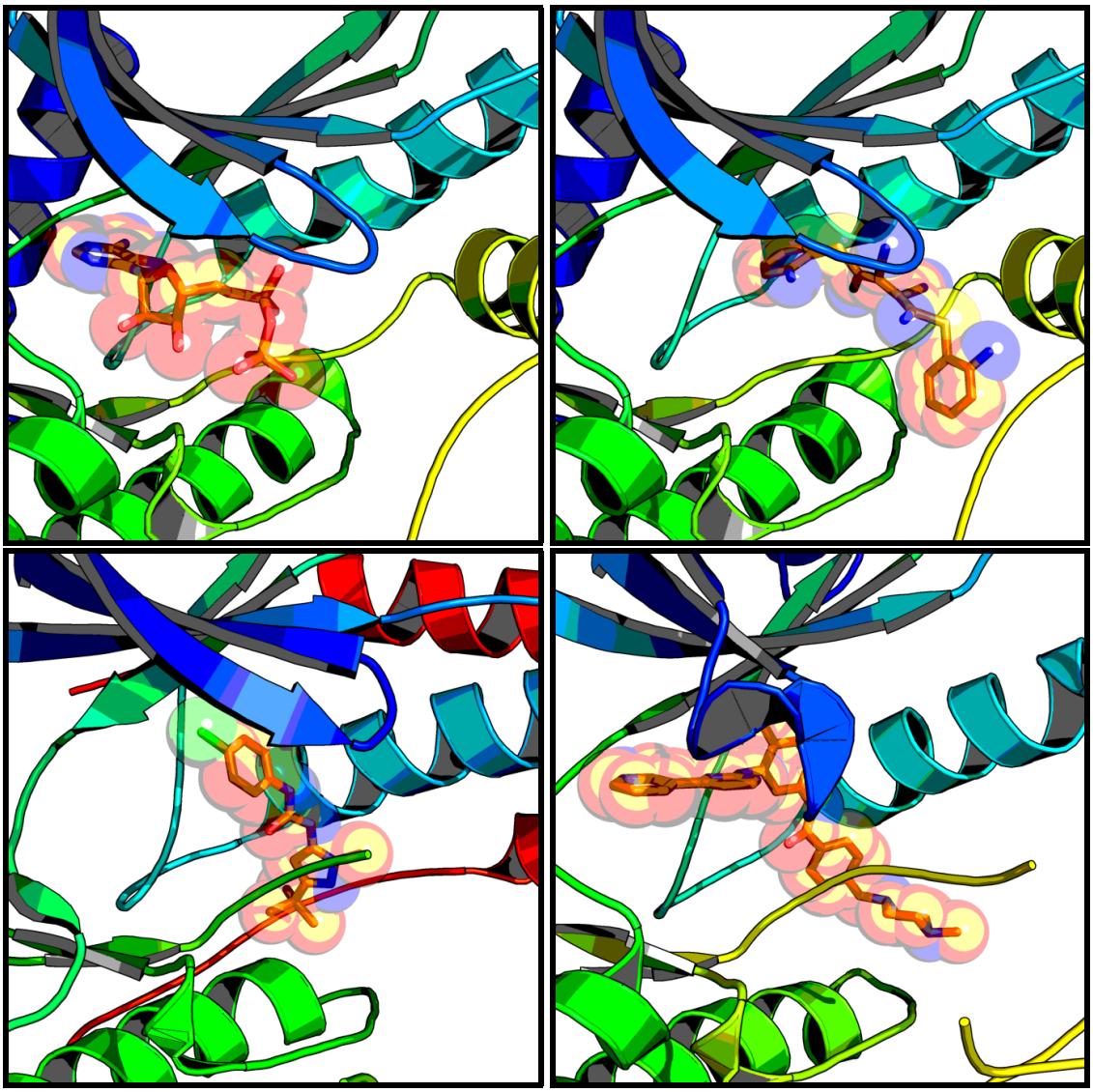


Figure 5.4: Four kinase ligands bound closely or overlapping the ATP binding site, that clustered into distinct groups in the hierarchical clustering. All protein structures oriented to the same reference structure. **Top left:** the ATP site, with ADP bound, **Top right:** non-ATP competitive allosteric site, with U0126 a *MEK1* inhibitor bound, **Bottom left:** ATP competitive allosteric site, with SKF86002 a *MAPK14* inhibitor bound and **Bottom right:** ATP competitive overlapping ATP site, with gleevec an *ABL2* inhibitor bound. PDB entry codes 3eqh, 3eqh, 1kv1 and 3gvu respectively. Figure generated using PyMOL (Delano, 2006).

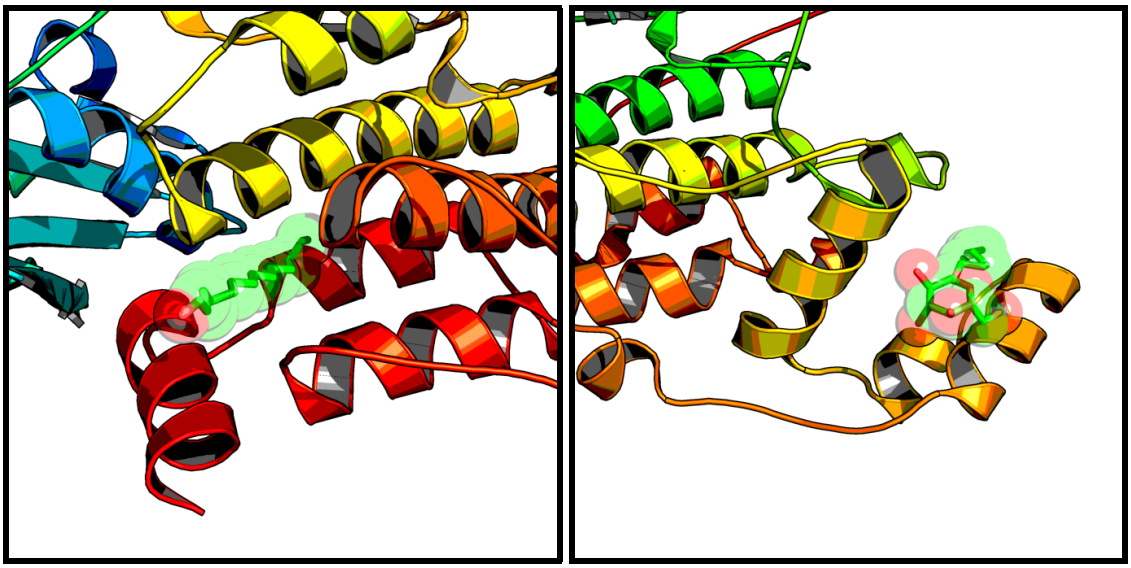


Figure 5.5: Two kinase ligands bound distantly to the ATP binding site, that clustered into distinct groups in the hierarchical clustering. Both protein structures oriented to the same reference structure. **Left:** the first C-terminal allosteric site, with myristic acid bound and **right:** the second C-terminal allosteric site, with β-octylglucoside bound. PDB entry codes 1opl and 2npq respectively. Figure generated using PyMOL (Delano, 2006).

tional variability in loops, crystallographic conditions, structural resolution errors and sequence alignment errors. However in most cases, their exists a core subset of residues which consistently interact with the ligand group, which enables clustering. To define the binding site positions, four methods were applied: **1. Loose binding site**, all positions in an identified cluster were added to the definition (e.g. Loose ATP site). **2. Conservative binding site**, any position which was observed in less than 35% of the cluster members was removed (e.g. Conservative ATP site). **3. Precise binding site**, each unique example of a ligand within the clusters was added as a binding site defined by its own structural contacts alone, if multiple examples of the same ligand were present, the example with the most contacts was chosen. **4. Whole domain binding site**, the kinase domain alignment was manually inspected and regions with large inserts in a small number of sequences were removed, leaving the core kinase domain regions and most loops.

5.2.9 Screening Data

In order to create knowledge based, predictive methods for ligand selectivity and polypharmacology, a set of ligands with known activity against our kinases was required. As described in Chapter 4, ChEMBL contains a large number of activities associated with protein kinases. While this is a potentially invaluable resource for future work, it currently has some limitations. As described in Chapter 4 much of this data is associated with multidomain proteins. So that any compound activity reported against a known kinase protein, may have arisen from an interaction with a non-kinase domain within the protein. A final problem which is

more subtle is that there may exist a bias in publication towards positive results, reducing the reporting of those screens where the compound was inactive, this has yet to be confirmed.

The major requirements for our test set were:

- A diverse set of compounds.
- A diverse set of protein kinases.
- Confidence that the activity of the compound is due to interaction with the kinase domain.
- Contains both active and inactive results for the same compound.
- Screening methods and activity units comparable.
- Compound structures available.

The pharmaceutical company, Abbott Laboratories recently collated its historical kinase screening data and made the data publicly available ((Metz *et al.*, 2011)). The data set comprised 3858 compounds, with activities reported against 172 human kinases. As the data were published as kinase enzyme inhibitory values, and the screening compounds were generally not specific to a single kinase (>96% of compounds showed activity against at least two kinases), there could be some confidence that the activity was due to interaction with the kinase domain. However, allosteric effects due to compound binding within a non-kinase domain could not be ruled out The set was not all-by-all (every compound screened against every kinase), but ∼40% of the matrix had screening data (Table 5.7). The inhibitory values were presented as pK_i, which is the negative $\log_{10}$ of the

K_i value. The K_i value is the binding affinity of the inhibitor and is directly comparable when compared across assays against different enzymes. Where the pK_i of a compound was low (i.e. low or no affinity), the values were expressed as less than the limit of the assay (e.g. pK_i <5.6), and could be interpreted for our purposes as inactives. Where the exact pK_i was presented, a pK_i cutoff determined if the compound was classed as active against the kinase. Of the assayed compounds ~40% were supplied with chemical structures, these were converted to Canonical SMILES (Weininger, 1988) string format using Pipeline-pilot (http://accelrys.com/products/pipeline-pilot/). Canonical SMILES strings were also generated for all CREDO ligands, so those Abbott compounds that were also in CREDO, could be identified by a database join. The kinase genes were supplied as gene names which were used to search for Uniprot identifiers to map to the kinase alignment. Six of the 172 genes contained dual kinase domains, and thus the inhibitory values against these genes could not be confidently assigned to one kinase domain so these genes were discarded from the analysis.

Abbott protein kinase compound screen	**Total**
Kinase genes assayed	172
Genes containing dual kinase domains	6
Compounds assayed	3,858
Compounds assayed with structures available	1,497
Activities assessed	258,094
Activities confirmed[1]	103,919

[1] where reported activity is not < limit of assay

Table 5.7: Abbott kinase screening file summary

Every pair of kinases in the Abbott screening set had pairwise sequence iden-

tity calculated over the residues in the **Loose ATP binding site** definition (48 positions). The binding affinities (pK_i) for all compounds that had been assayed against both pairs were correlated using the coefficient of determination (R^2). Four selected pairs of kinases are shown in Figure 5.6. Panel (A) shows a pair of kinases with a high sequence identity over the binding site, as would be expected, the correlation between compound activities here is strong ($R^2 = 0.81$). Panel (B) shows a pair of kinases with a lower sequence identity over the binding site, but a stronger correlation of activities ($R^2 = 0.89$). It should be noted that this is possibly due to far fewer data points. Figure 5.6 (C) and (D) each show a pair of kinases with 45% sequence identity over the binding site, but with very different correlation of activities (R^2 0.48 *vs.* 0.00). An overview for all pairs of kinases is shown in Figure 5.7. The drop in correlation is sharp after around 80% identity over the binding site, but even at these levels and greater, there exists a high standard deviation in the correlations.

5.2.10 Compound Binding Inference

Two methods are described here to attempt to infer compound binding from a training set of kinases, onto a test set of kinases.

5.2.10.1 Amino Acid Conservation Model

A simple method to predict if a compound will inhibit a protein, is to infer from evidence that the compound inhibits a closely related protein. It is logical to assume that if the determinants of ligand binding (i.e. the amino acid residues which form the ligand binding pocket), are identical between two proteins, then the ligand binding profile should be the same (excluding any allosteric effects).

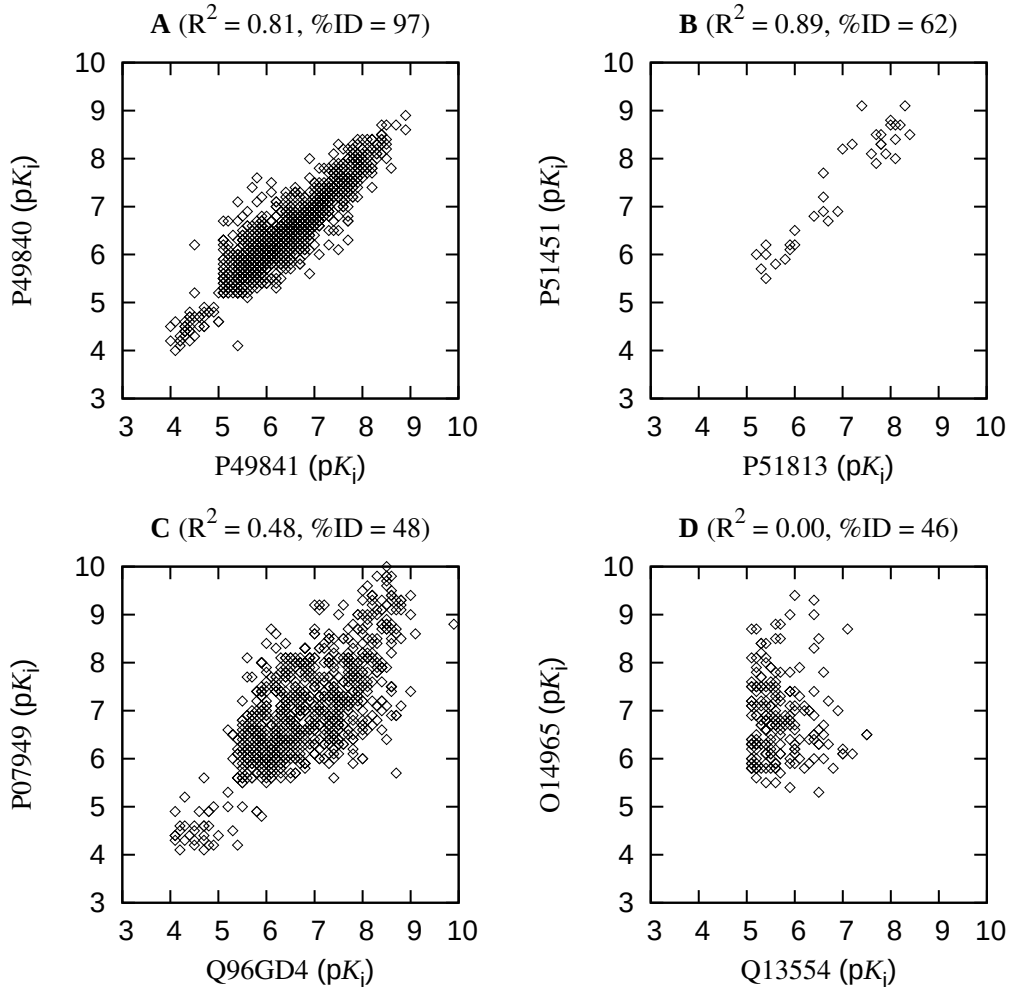


Figure 5.6: Correlation of ligand binding profiles in highly similar binding sites (A, B) and less similar sites (C, D). Each point represents an Abbott compound and its activity (pK_i) against a pair of kinases. The kinases are denoted by their UniProt entry accession. The correlation of activities is measured using R^2 (coefficient of determination). The sequence identity (%ID) of the kinase pair is calculated over the 48 residues in the **Loose ATP binding site**.

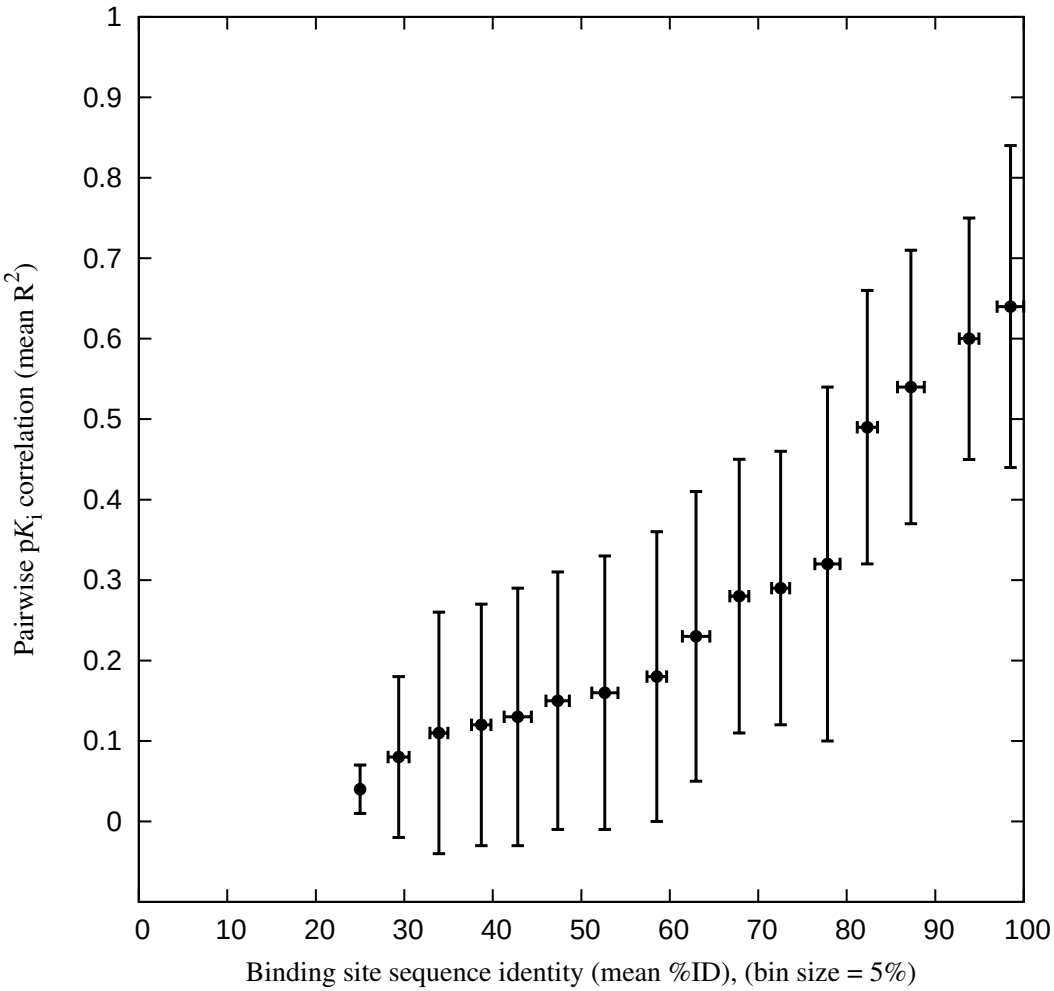


Figure 5.7: Correlation of Abbott compound-kinase activities versus binding site sequence identity. The sequence identity (%ID) of the kinase pairs is calculated over the 48 residues in the **Loose ATP binding site**. Kinase pairs are binned by sequence identity, and the mean value shown for each bin. The mean pK_i correlation of activities is shown, for kinase pairs in each sequence identity bin. Bars show standard deviation. (The underlying distribution of binding site sequence identity is shown in Appendix B. Figure A.4.)

While in general kinases are much more conserved at ligand binding positions than over the whole domain, in practise few proteins share 100% identity over the binding site. By applying a variable cutoff, the activity of a compound against a kinase, could be inferred if another kinase within the identity threshold was inhibited by the same compound.

5.2.10.2 Naïve Bayesian Model

Bayes' Theorem is a way of inverting a conditional probability. It is stated:

$$P(X|Y) = \frac{P(Y|X)P(X))}{P(Y)} \tag{5.2}$$

which means the probability event X being true given event Y has occurred. It can be used in a predictive manor by substituting X with a category (e.g active compound), and substituting Y with a feature of the protein (e.g. hydrophobic residue at position 24 of the alignment). Given a training set of examples of compounds and the protein feature, its is possible to calculate the P(Y|X) as the probability of feature present given an active compound, P(X) as the probability of active compound independent of the feature, and P(Y) as the probability of the feature present independent of the compound being active:

$$P(active|feature) = \frac{P(feature|active)P(active))}{P(feature)} \tag{5.3}$$

The same can be calculated for the compound being inactive:

$$P(inactive|feature) = \frac{P(feature|inactive)P(inactive))}{P(feature)} \tag{5.4}$$

With both categories being calculated, a prediction would be based on which probability was greatest such that:

$$active = P(active|feature) > P(inactive|feature) \tag{5.5}$$

$$inactive = P(inactive|feature) > P(active|feature) \tag{5.6}$$

Where multiple features are available, the probability of each can be multiplied to calculate the overall probability of the event given all features. As this method assumes the features are independent of each other (i.e. state of one feature does not affect the state of another feature) it is called the "Naïve Bayesian". In this case, the features of a protein are not independent. However, it has been shown that making this independence assumption even when it is not true, often has little effect on the results (Domingos & Pazzani, 1997).

5.2.10.3 Naïve Bayesian Implementation

The Naïve Bayesian Models were implemented in Perl using the Algorithm:: NaiveBayes module from CPAN (http://www.cpan.org/). The data was divided into Active and Inactive target-compound pairs using a pK_i cutoff of >6. Bayesian models were created using a defined binding site (site as described by the kinase MSA), and a set of AAindex descriptors (Figure. 5.8). The weights for the features were the normalized AAindex values for the amino acid at that alignment position for the kinase training set (Figure 5.9). The Bayesian implementation software used could not utilize negative values for features, so the normalized AAindex value range used was [1 .. 10]. Multiple proteins with activ-

ity data against the compound could be used to train the Bayesian model, which in turn could be used to predict the compounds activity against a protein with no screening data (Figure 5.10). Each model could be trained on any number of examples, using multiple binding site positions and multiple AAindex descriptors with no significant performance issues.

Figure 5.8: **Describing a Bayesian model**. There are two possible model states for a compound, active or in-active against a protein. The proposed ligand binding site is defined by a set of multiple sequence alignment (MSA) positions, in this case four positions of the MSA. The ligand binding site residues are described by amino acid property descriptors, in this case: blue properties described volume (GRAR740103) and orange described polarity (GRAR740102).

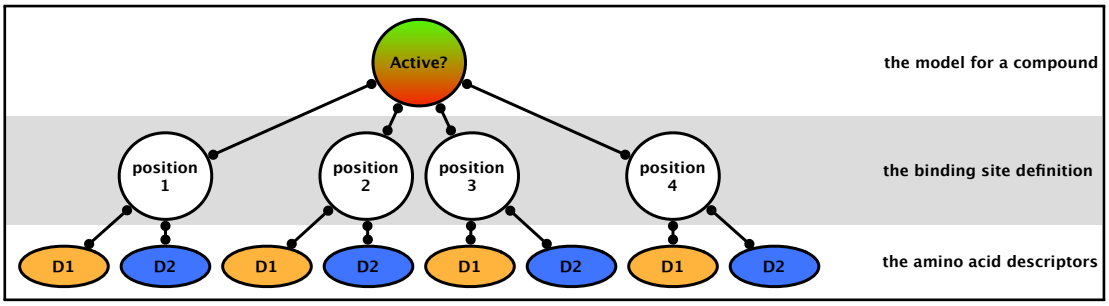


Figure 5.9: **Populating a Bayesian model**. In this case, the compound is potent (green) against protein *CDK2*. The residues at the MSA positions for *CDK2* are added, shown here colored by the ClustalX scheme (Thompson, 1997). The residue properties are added for each descriptor, the shade of each property relates to the size of the property descriptor.

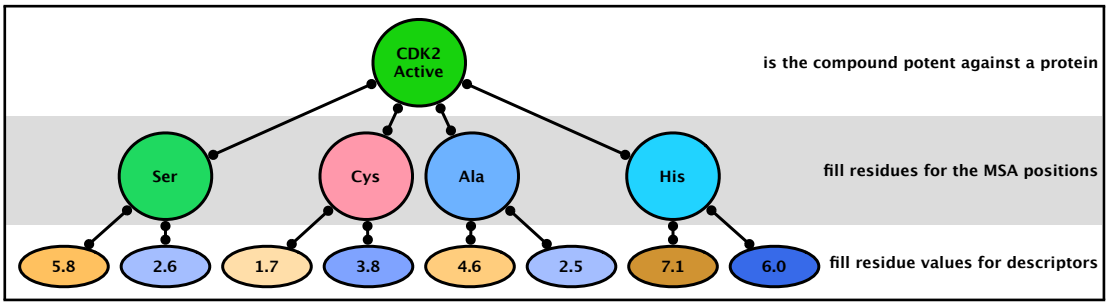
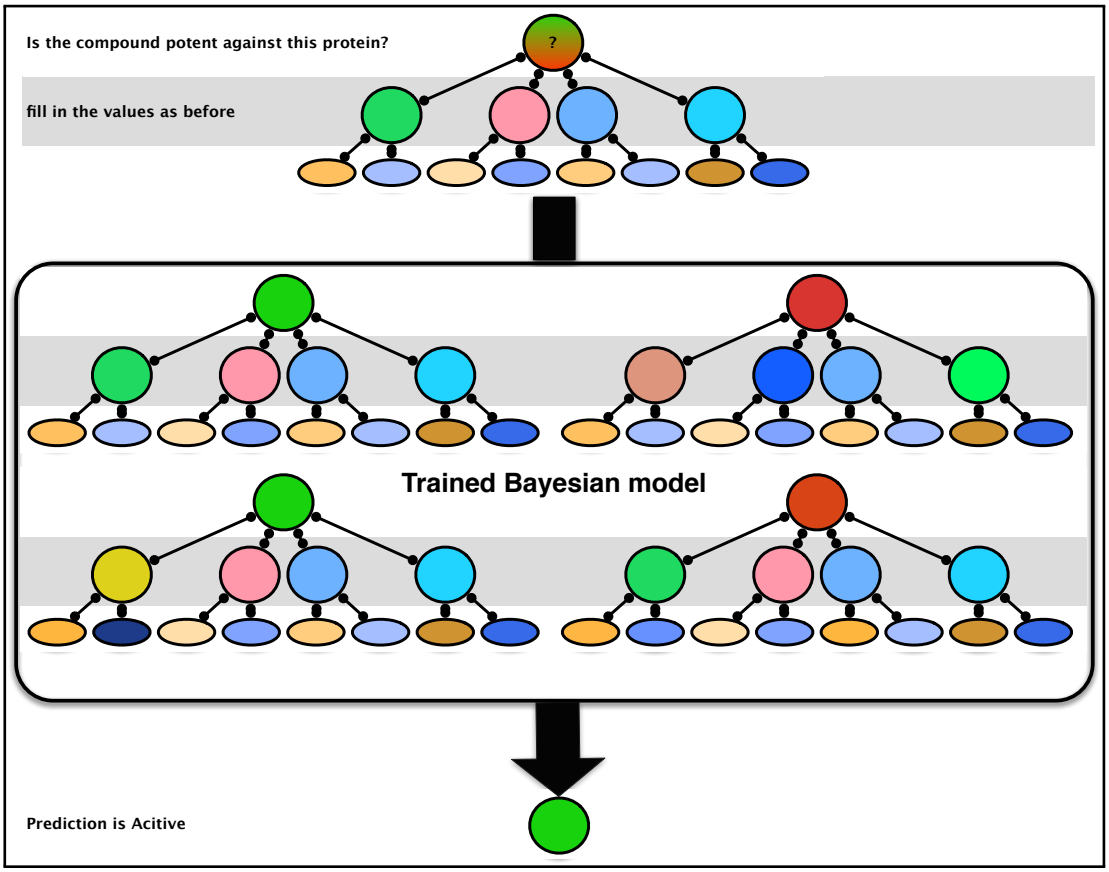


Figure 5.10: **Classifying with a Bayesian model**. In this case, the potency of the compound is against the protein is unknown (shaded green to red). However the residues at the MSA positions are known, as are the descriptors. These are filled as in Figure 5.9. Multiple training proteins which are inhibited (green) or not (red) by the compound, are populated as before. The Bayesian model then predicts the potency state for the compound against the unknown protein.



5.2.10.4 Validation Sets

In order to assess the validity of the models for transferring activity, the data set could not wholly be used to train the models, as some data is required to be retained in order to benchmark the models. Another question to ask of the models is how many known data points are required to train them, and how it affects the accuracy of the predictions. The training set could also be biased by aberrant kinases, those which are either promiscuous compound binders or super-selective could affect the benchmark. To combat this, the data set was divided into training and testing sets randomly, and 50 iterations of the benchmark performed so most of the data points would be used to train and test independently. Where a compound had been assayed against less than 70 of the kinases, these compounds were discarded from the benchmark. The number of training examples was varied between 5 and 60, and the remainder used to test the model. For each model, each of the 50 training and testing sets were used independently, and the average prediction accuracy calculated. All of the variations of prediction models used the same 50 training and testing sets so they could be compared legitimately.

5.3 Results

5.3.1 Benchmark of the binding site identity models

The results of the benchmark of using percent identity of residues in the binding site to infer shared activity are shown in Figure 5.11. As can clearly be seen, a stringent identity threshold (90%) while producing very few false positives (incorrectly predicted actives), also misses most of the true positives. At 50% identity

across the binding site, the inference even when using a large set of training compounds, is not much better than a random guess. Where larger amounts of training data are available (40+), a sequence identity cutoff of 60% could be used to infer potential polypharmacology targets, as with a large family of proteins such as the kinases, the TPR observed at these levels (≈ 0.48) would provide a substantial number of targets to investigate. However, this method would be less useful for predicting selectivity, as many proteins that would share compound binding profiles would not be inferred.

5.3.2 Benchmark of the Bayesian models

Figure 5.12 shows the benchmark using different definitions of the binding site residues, and how they affect the Bayesian models predictive power. Only those compounds that were in both the Abbott screening data, the validation set (5.2.10.4) and observed to be kinase-bound in CREDO were analyzed. The *Sheinerman* descriptors were used to describe the binding site properties. With low numbers of training examples, all these definitions perform poorly (PC_d of between 0.95-0.91 for 5; and 0.95-0.89 for 10 training examples). As more training data are utilized, each model's sensitivity improves at a similar rate.

Figure 5.13 shows the benchmark using different descriptors of the binding site properties, and how they affect the Bayesian model's predictive power. The **Loose ATP binding site** definition was used to describe the binding site residues. All compounds in the validation set (5.2.10.4) were used. Increasing the training data increased the sensitivity for both sets of descriptors, with a slight cost to specificity.

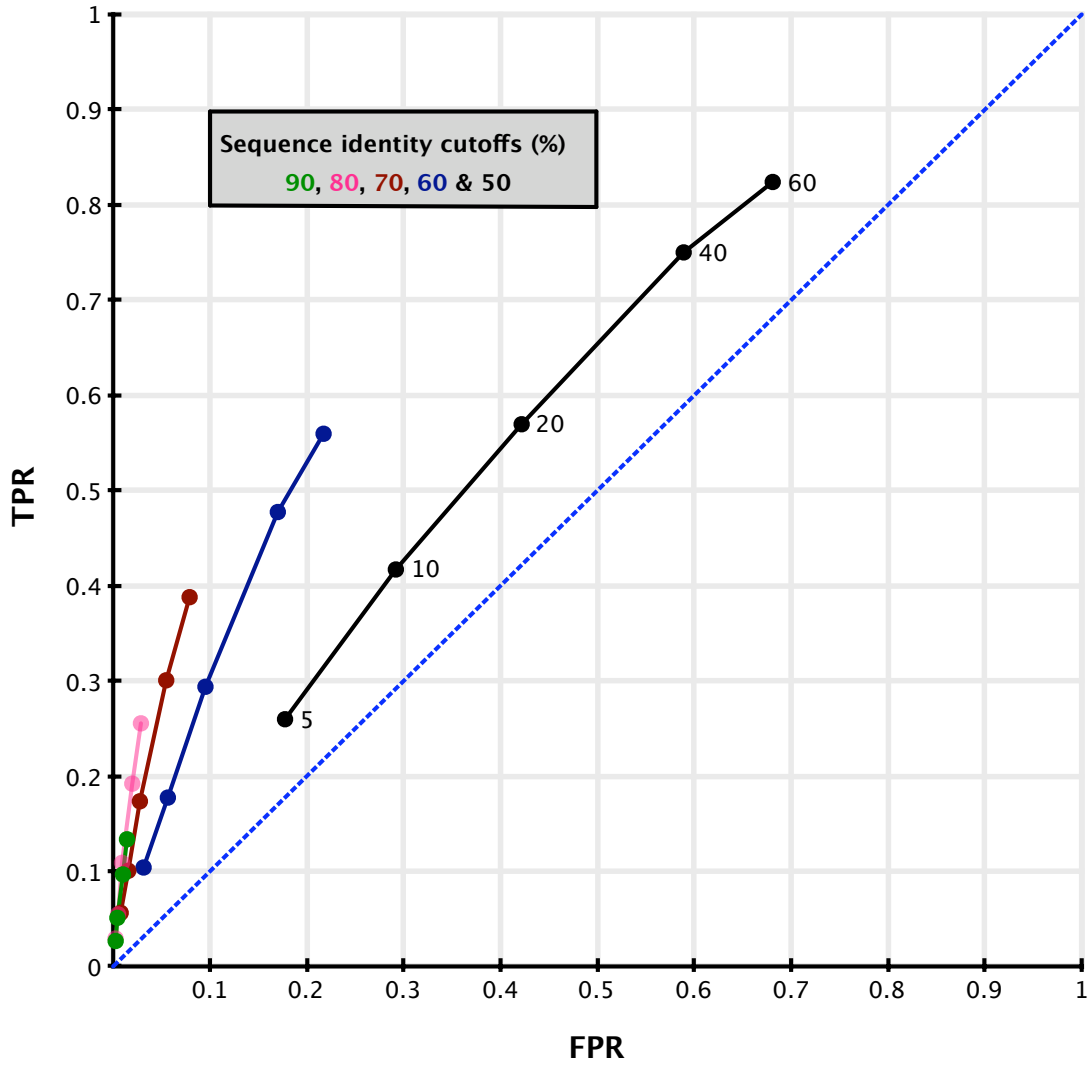


Figure 5.11: ROC analysis of the Amino acid conservation model as described in 5.2.10.1. The analysis was performed on all compounds in the Screening data. Sequence identity calculations were performed on the residues in the **Loose ATP binding site** definition. Number of training data points are shown in black.

The **Loose ATP binding site** definition of the binding residues, using the *Sheinerman* binding site property descriptors, was common in both benchmarks, the only difference between the two was the set of compounds used to benchmark the Bayesian. The compounds used in the **Ligand binding site properties benchmark** (Figure 5.13) entirely subsumes the compounds used in the **Ligand binding site residues benchmark** (Figure 5.12), so the difference in performance can only be attributed to a random bias of poorly performing compounds in the smaller set.

As with the binding site identity predictions (Figure 5.11), the predictions based on the Bayesian model (Figure 5.12) are much more specific (low FPR) than sensitive (high TPR). In practical terms, when a compound has been observed to bind potently to a set of proteins, the compound can be predicted to bind potently to other proteins of the family, and those predictions are likely to be correct. Conversely, where the compound is predicted not to bind a protein, a large number of these proteins would in reality be inhibited by the compound. This would cause problems when trying to assess selectivity, as many non-selective compounds could be predicted as selective.

5.4 Conclusions and future direction

The structure-ligand information available from CREDO (Schreyer & Blundell, 2009) provides an invaluable resource for linking structural contact information to binding site residues, and through the process described here, onto all domain family members. This simple process can be applied to any protein family of interest - given any structural representation for the domain family - to rapidly

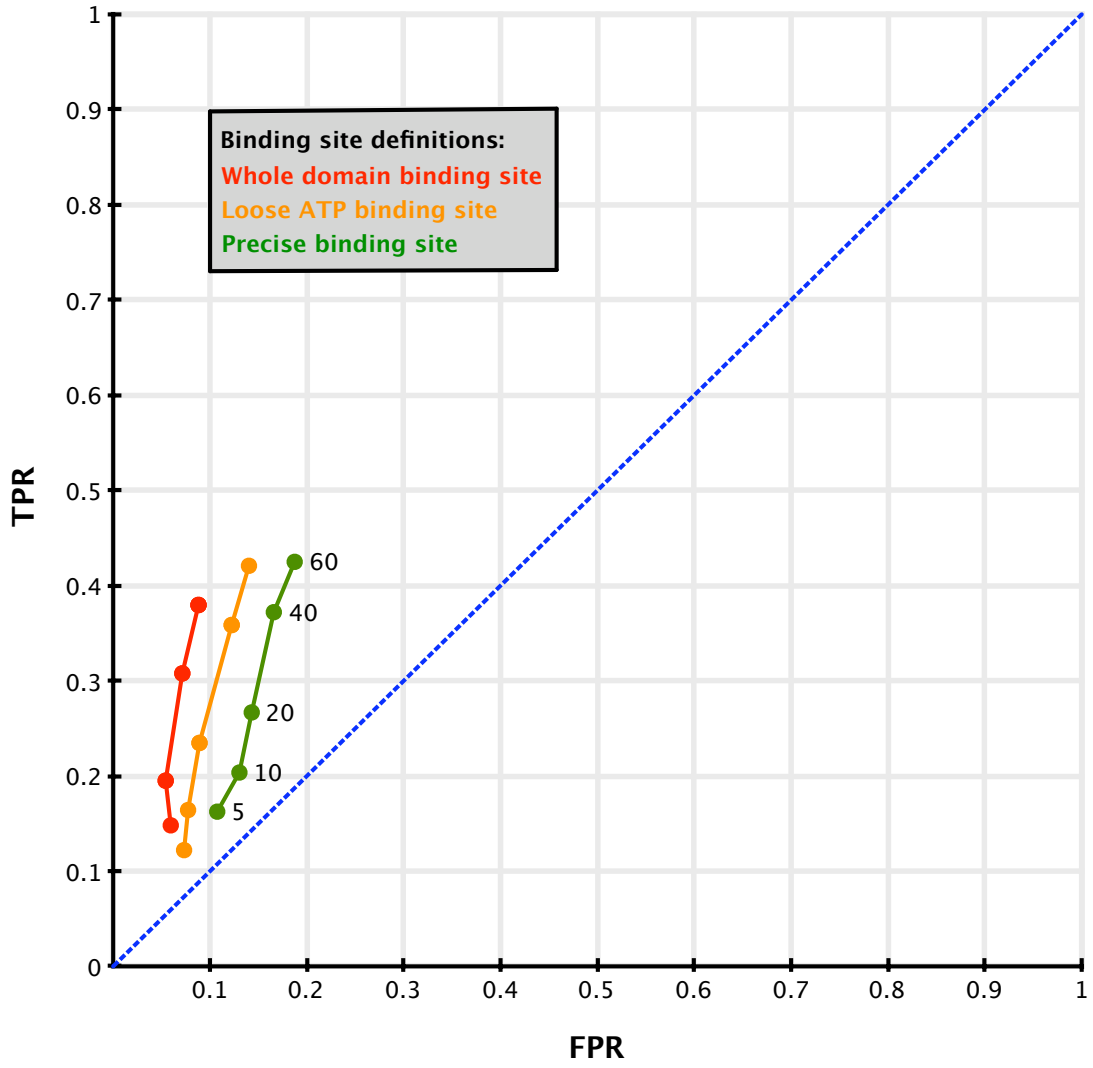


Figure 5.12: **Ligand binding site residues benchmark**. ROC analysis of the effect of the ligand binding site definition on the Bayesian inference model (5.2.10.2). The analysis was performed on just the compounds in both the Screening data (5.2.9) and kinase-bound in the structure database (5.2.3). Trends for the **Whole domain binding site** are shown in red, **Loose ATP binding site** in orange. The **Precise binding site** (structurally observed ligand binding positions for each compound) in the green series. The number of training kinases shown in black.

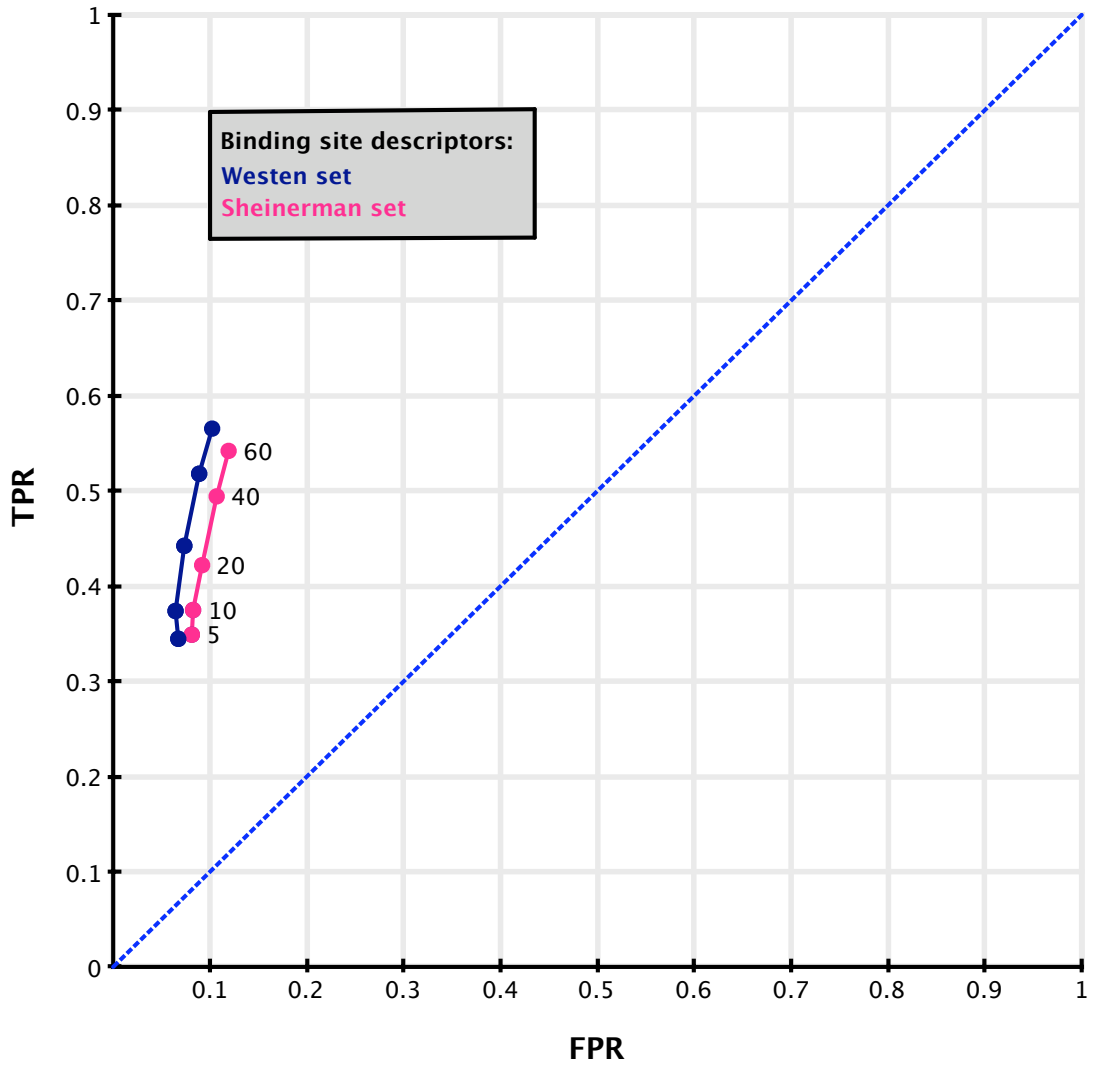


Figure 5.13: **Ligand binding site properties benchmark**. ROC analysis of the effect of increased binding site feature information on the Bayesian inference model (5.2.10.2). The analysis was performed on all compounds in the Screening data. Identity calculations were performed on the **Loose ATP binding site**. Green is *Westen* descriptors. Red is *Sheinerman* descriptors, . The number of training data points shown in black

catalogue potential drug binding sites.

These ligand binding sites can be used to infer selectivity across a target family. A clear caveat on such analysis is that even between highly similar binding sites there can be large variations in ligand binding profiles.

The Bayesian classifier predicts whether a compound will be active against a set of proteins, using related protein screening results as a model. Unsurprisingly, larger amounts of training data produce better results. This could be an issue for lesser studied or smaller, protein families that lack significant amounts of suitable and accessible bioactivity data.

These results suggest that knowledge of the specific binding mode of a compound does little to improve inference, as using the most precise definition of a binding site increases the sensitivity of predictions marginally, accompanied by a larger decrease in specificity. Adding more information to describe the properties of the binding sites does appear to improve the models, albeit only marginally.

When applying these approaches it is important to bear in mind that the lack of sensitivity means that potential selectivity issues may be missed. However, given the levels of specificity observed in Figure 5.2.10.2 broad spectrum targets or polypharmacology targets that are identified with this method, will be enriched with many more truly non-selective proteins than selective proteins.

Chapter 6

Conclusions

6.1 Overview

The primary objectives of the research presented here was to address three distinct criteria of pathogen target selection. To date, the essentiality prediction can be used, and have been employed to prioritize potentially essential proteins from a genome. The druggablity module, can be used "as is" to search for druggable domain profiles within a proteome. The caveat being it will not be as sensitive as a domain-based model. The selectivity analysis relies heavily on domain family specific screening data, and to date only the protein kinase family has been investigated. However, as the protein kinases are a large druggable family present in many eukaryotic pathogens, the analysis and prediction of potential pathogen kinase inhibitors is feasible.

6.2 Essentiality

Many existing anti-infectives target multiple non-essential proteins simultaneously to produce a lethal response. Currently, our ability to elucidate these

inter-dependent targets is limited. Despite the relatively small genomes of many pathogens, the combinatorics of finding just the gene pairs that are essential in tandem would require genome-size^2 targeted gene deletions. Given this caveat, single gene essentiality is still an important criterion for target selection, and many current antibiotics exploit this. For most pathogen genomes we do not have genome-scale experimental essentiality data, and despite a recent surge in the number of whole-genome essentiality screens (Christen *et al.*, 2011; Klein *et al.*, 2012; Xu *et al.*, 2011), the rate of experimentally derived essentials genomes is too slow for the systematic application to all pathogens. Therefore, *in silico* genome essentiality prediction methods are required that capitalize on existing genome scale essentiality experiments. Previously to this work, the idea of using homology and orthology to known essentials, as indicators of essentiality had been employed (Aguero *et al.*, 2008; Holman *et al.*, 2009), but without any insight into how effectively they worked across multiple species. Chapter 2 attempted to address this, with the development and benchmarking of a procedure to harnesses the available experimental data on genetic essentiality and uses a series of phylogenetic models to infer likely essentiality in related pathogen proteins.

To date, data on published genome scale essentiality experiments have been incorporated into a scalable relational database framework, which enabled the application a robust orthology detection algorithm at large scale. Identifying orthologs and paralogs shared between experimentally characterized organisms and pathogenic genomes of interest, enables putatively essential genes to be identified by inference. The application of multiple models of essentiality allowed a ranking of putative essentials rather than a binary classification, which is important given that the most specific models of essentiality suffer from a low recall rate. Simul-

taneously to my work benchmarking the method in bacteria, Doyle *et al.* (2010) performed a similar benchmark on eukaryotes. Comparison of both results, show that orthology and orthology to a known essential gene are positive predictors of essentially in both eukaryotes and prokaryotes.

Since this work, Yuan *et al.* (2012) have used a machine learning method trained on multiple *Ab initio* calculable protein properties, and identified the gene evolutionary age, as the top predictive feature. This property is comparable to my most selective model, as multiple orthologs across many species greatly increased the chances of being essential, and where these relationships still exist, then the likelihood is that their conservation throughout speciation was essential.

In Chapter 3, these procedures were applied to three different anti-infective discovery projects, being undertaken at the University of Dundee and the London School of Hygiene and Tropical Medicine. These include the genomes of the Gram negative bacterium *Pseudomonas aeroginosa*, a panel of seven kinetoplastid genomes and the genome of the clinically important parasite *Schistosoma mansoni*. The analysis of essentials predicted in *S. mansoni* represented a recurring theme: at some (undetermined)-evolutionary distance, the types of genes predicted essential tended to be involved in "core" metabolism. A clear disadvantage of this are the potential selectivity issues with the host, and more importantly, only a small number of predicted essentials. These limited sets can quickly be exhausted when applying further filters such as druggability.

The question - "which of these non-essential proteins can i infer essentiality onto"? Is most likely the cause of the lack of sensitivity in my own, and others methods. A logical next step would be to ask the question - "Assuming all proteins are essential, can i find evidence to infer them non-essential?".

6.3 Druggability

Druggability is an equally important criteria for assessing potential drug targets alongside essentiality, particularly given the failure of large-scale, high-throughput screening in discovering new lead compounds against antibacterial targets (Payne *et al.*, 2006). Multiple methods exist to predict druggability, either based on structural features (Halgren, 2009), precedence (Hopkins & Groom, 2002) or using calculable protein features (Al-Lazikani *et al.*, 2007). For any validation, all methods require testing data, proteins that are known to bind drug-like molecules with high affinity. The ChEMBL database offers a window into this valuable training data, but how valuable is the data in it current form?

In Chapter 4, the protein-ligand information available from ChEMBL was assessed for its validity as a resource for inferring the druggability of pathogen targets. What was clear, were the challenges faced to fully harness the data to mirror biological reality. A major issue with the early ChEMBL version, was the linkage of compounds via affinity to the protein sequence level. In reality, the association of a compound to a protein is via a specific set of residue-ligand interactions. A simplification of this complex reality is to assign the ligand to a protein domain or set of domains that can be defined as a "druggable unit". The programmatic procedure described in Chapter 4 was a first step in defining the potential druggable units. The annotation provided a "domain fingerprint", which combined with a simple measure of associated compound desirability, enables other proteins with the same domain organization to be inferred as druggable.

Since this work was undertaken, the ChEMBL team have implemented their own system of domain annotation (Kruger *et al.*, 2012), based on PFAM. Using

these domains they attempted to automatically assign compounds to domains using a simple heuristic of "seed domains". The seed domains are those which occur as a single domain in ChEMBL, and where they also occur in a multidomain, the corresponding "seed domain" is assigned as the binding domain. With this approach they predict the correct domain will be assigned in 88% of multidomain cases. It is clearly a large task linking domains and compounds in the large tail of difficult edge-cases, and ultimately hand curation of these cases may be necessary.

The applicability of ChEMBL to bacteria was briefly discussed, given the lack of prokaryotic targets in ChEMBLv01. Since this analysis, the database has grown rapidly, from 440,000 compound screened against 3,622 protein targets, to 1.3 million against 6,235. However, the proportion of eukaryotic and prokaryotic targets has remained broadly the same.

6.4 Selectivity

The ability of some compounds to bind multiple members of a protein family is based not only on the compounds properties, but also on the properties of the protein's binding site. Where a pair of proteins exhibit similar physicochemical properties at positions important for compound binding then they may share compound binding profiles.

In Chapter 5, the ability to predict selectivity or non-selective proteins was assessed using the protein kinase domain family as a test case. The protein kinases are a large and diverse family, and exhibit multiple ligand binding modes. The availability of large scale compound binding affinities was a major factor

in selecting the test case. For the proteins kinases, the Abbott kinase screening data provided a large compound-kinase screening matrix, with both active kinase-compound pairs and inactive pairs.

The analysis of the structural information in the CREDO database, enabled the semi-automated classification of distinct ligand binding modes, including allosteric sites. Importantly, limiting a selectivity assessment to a single domain family, allows the transfer of structural information across the whole family, using a multiple sequence alignment.

The classical selectivity assessment is to compare the sequence identity of the binding site residues in a pair of proteins, and where the identity is high, then these two proteins would be considered non-selective and would share similar ligand binding profiles. In order to establish the validity of this assessment, a benchmark was performed. The results were unsurprising, protein binding sites with high sequence identity often share compound binding profiles, however, protein binding sites with low sequence identity often share compound binding profiles also. There are two reasons assumed for this, firstly, sites can be very similar in terms of physicochemical properties, but share low sequence identity due to redundant properties of the 20 standard amino acids, and secondly, specific modes of compound binding can be entirety dependent on very small number of binding site residues, and while overall the binding site is not conserved, the important residues could be identical.

To address the limitation of the sequence identity method, a machine learning method using a Naïve Bayesian classifier to learn these patterns of amino acid properties, a used them to predict compound binding. Unsurprisingly, where amounts of training data produce better predictions. This could be an issue

for lesser studied protein families that lack significant amounts of suitable and accessible bioactivity data. These results suggest that knowledge of the specific binding mode of a compound does little to improve inference. Adding more information to describe the properties of the binding sites does appear to improve the models, albeit only marginally.

When applying these approaches it is important to bear in mind that the lack of sensitivity means that many potential selectivity issues may be missed. However, given the good levels of specificity broad spectrum targets or polypharmacology targets can be identified with reasonable confidence. This sensitivity/selectivity bias, indicates that rather than a selectivity predictor, i have inadvertently produced a promiscuity predictor.

Little comparable work has been done in this field, while medicinal chemists often use specific knowledge of binging site properties to guide compound design, it is not often in an automated fashion. However, they use properties of the compound also, which is lacking from this method. Another feature of the data currently unused it the cross-correlation of binding profiles, where the knowledge that a compound is potent against a protein, may infer that similar compounds are more likely to be potent against the same protein.

A logical next would be to analyze a multiple different protein families, to see if the kinases are typical of performance or a difficult case.

6.5 Outlook

The modular informatics framework presented here could be a useful set of methods to improve the exploitation of genomic information in anti-infectives drug

discovery, by enabling proteome-based, domain-based and binding site based comparisons within and between genomes.

Appendix A

Appendix

A.1 Proteome database

Table A.1: The database tables for proteome, essentiality and orthology information. The datatypes for each column are shown. Each database tables primary key (where available) is shown (**pk**). The reference table(s) that post foreign keys are shown in **fk table**. The **notes** section described the column data with example where appropriate.

Proteomes - describes a published proteome (or the protein compliment of a genome study).				
column name	**type**	**pk**	**fk table**	**notes**
Proteome_id	Number	Y		Internal proteome identifier
Proteome_reference	String		Newt_entry	Description of the source paper or online resource of the the original genome or proteome information.
Load_date	Date			Date proteome added to the database (**not** the date of proteome release).
Taxon	Number			NEWT taxon identifier for proteomes species
Proteins - describes the proteins of a proteome.				
column name	**type**	**pk**	**fk table**	**notes**
Protein_id	Number	Y		Internal protein identifier
Proteome_id	Number		Proteomes	Internal proteome identifier
Accession	String			The unique per-proteome accession used by the original proteome resource. e.g. "ACIAD1303"
Gi_number	Number			The NCBI sequence identifier (where available).
Primary_description	String			The protein description provided by the original proteome resource. e.g. "putative extracellular nuclease"
AA_sequence_digest	String			The md5 checksum/digest of the proteins ammino acid sequence.
AA_sequence	Large			The proteins ammino acid sequence.
Protein_attributes - describes the diverse additional annotation of proteins.				
column name	**type**	**pk**	**fk table**	**notes**
Attribute_id	Number	Y	Proteins	Internal attribute identifier.

Continued on next page

Table A.1 – *Continued from previous page*

Attribute_value	String			Other unrequired data value provided with a protein annotation e.g. "*gapB*".
Attribute_name	String			Other unrequired data type provided with a protein annotation e.g. "gene name".
Protein_id	Number			Internal protein identifier.
Proteome_orthomcl - describes the co-orthology relationships between a pair of proteomes, as calculated by orthomcl.				
column name	**type**	**pk**	**fk table**	**notes**
Ortholog_id	String	Y		Internal co-ortholog relationship identifier.
Proteome_id_a	Number		Proteins	Internal proteome identifier for species A.
Proteome_id_b	Number		Proteins	Internal proteome identifier for species B.
Accession_a	String		Proteins	The protein accession of species A.
Accession_b	String		Proteins	The protein accession of species B.
Cluster_id	Number			Internal identifier for a co-orthologos cluster.
NEWT_entry - describes the EBI NEWT taxonomy hierarchy.				
column name	**type**	**pk**	**fk table**	**notes**
Taxon	Number	Y		NEWT taxon identifier for this group
Scientific_name	String			Name of taxonomic group e.g. "*Homo sapiens*"
Rank	String			Taxonomic rank e.g. "Species"
Parent_taxon	Number		Newt_entry	NEWT taxon identifier for this groups parent group
Essential_sets - describes the published essentiallity screens.				
column name	**type**	**pk**	**fk table**	**notes**
Set_id	Number	Y		Internal identifier for this essential screen.
Taxon	Number		Newt_entry	NEWT taxon identifier for essential screen species.
Literature_ref	String			Citation for the paper describing the screen.

Continued on next page

Table A.1 – *Continued from previous page*

Genome_wide	Number			Flag for whole geneome essentiallity screen.
Essential_genes - the essential proteins from published essentiallity screens.				
column name	**type**	**pk**	**fk table**	**notes**
Accession	String			The proteome accession used by the original proteome resource.
Set_id	Number		essential_sets	Internal identifier for this essential screen.
DEG - describes the DEG (database of essential genes) resource.				
column name	**type**	**pk**	**fk table**	**notes**
DEG_accession	String	Y		DEG identifier
GI_number	Number			The NCBI sequence identifier.
Taxon	Number			NEWT taxon identifier of the essential screen species.
Description	String			The protein description provided DEG.
AA_sequence_digest	String			The md5 checksum/digest of the proteins ammino acid sequence.
AA_sequence	Large			The proteins ammino acid sequence.

A.1.0.1 Proteomes database usage examples

Use orthomcl results to predicted essential proteins in *Pseudomonas aeruginosa* PAO1, using orthology to known essential genes in *Escherichia coli*. This query produces cluster_ids which represent a co-orthologous group between the two species. If there are multiple proteins from *P. aeruginosa* PA01 in this cluster then those protein had in-paralogs, and therefore represented a model m3 prediction (see section 2.2.7). Single *P. aeruginosa* PA01 proteins represent model m4 predictions.

```
SELECT proteome_id_a,
       proteome_id_b,
       cluster_id,
       -- how many predicted proteins in co-ortholog group
       COUNT(DISTINCT prt_a.accession) total,
       -- if more than one predicted protein has in-parlaogs
       DECODE(COUNT(distinct prt_a.accession),
              1, 'ORTHOLOG',
```

```
                'PARALOGS'
      ) in_para
  FROM proteome_orthomcl mcl
  JOIN proteomes pa
    ON pa.proteome_id = mcl.proteome_id_a
  JOIN proteomes pb
    ON pb.proteome_id = mcl.proteome_id_b
  JOIN proteins prt_a
    ON mcl.accession_a = prt_a.accession
  JOIN proteins prt_b
    ON mcl.accession_b = prt_b.accession
  JOIN essential_sets es
    ON es.taxon       = pb.taxon
  JOIN essential_genes eg
    ON eg.set_id      = es.set_id
-- want to predict essential co-ortholog clusters in this
   species
 WHERE pa.taxon = (
        SELECT taxon
          FROM newt_entry n
         WHERE n.scientific_name = 'Pseudomonas aeruginosa
            PAO1')
-- using this species' known essential genes to predict
   AND pb.taxon = (
        SELECT taxon
          FROM newt_entry n
         WHERE n.scientific_name = 'Escherichia coli str. K-12
             substr. MG1655')
   AND eg.accession = prt_b.accession
 GROUP BY proteome_id_a, proteome_id_b,cluster_id;
```

The first 6 results of the query are shown below, there are two clusters of m3 predictions, and four m4 predictions:

```
PROTEOME_ID_A |PROTEOME_ID_B |CLUSTER_ID |TOTAL |IN_PARA
-------------------------------------------------------------
3             |2             |53         |2     |PARALOGS
3             |2             |1300       |1     |SINGLE
3             |2             |1700       |1     |SINGLE
3             |2             |1900       |1     |SINGLE
```

```
3              |2              |2100        |1      |SINGLE
3              |2              |253         |2      |PARALOGS
---------------------- truncated -----------------------------
```

The following SQL inspects the first cluster shown above:

```
SELECT DISTINCT
       prt.protein_id,
       mcl.accession_a,
       prt.primary_description
  FROM proteome_orthomcl mcl
  JOIN proteins prt
    ON prt.accession      = mcl.accession_a
 WHERE prt.proteome_id    = mcl.proteome_id_a
   AND mcl.cluster_id     = 53
   AND mcl.proteome_id_a = 3
   AND mcl.proteome_id_b = 2;
```

Which finds the two proteins in the m3 cluster (both probable serine proteases):

```
ACCESSION_A |PRIMARY_DESCRIPTION
-------------------------------------------
PA0766      |serine protease MucD precursor
PA4446      |AlgW protein
```

A.2 Homology inferrence

Table A.2: Homology benchmark details. Performance of predicting essential proteins in five species using homology to DEG essentials, or homology to four essential proteomes. (where **Cov.** = BLAST alignment coverage (%) of the known essential protein, TPR = true positive rate, FPR = false positive rate, PPV = positve predictive value, $\overline{TPR}$ = average TPR, $\overline{FPR}$ = averaged FPR and $\overline{PPV}$) = average PPV.

		***Vs.* DEG**						***Vs.* 4 proteomes**					
Cov.	**Predicted species**	**TPR**	**FPR**	**PPV**	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$	**TPR**	**FPR**	**PPV**	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$
95	*Acinetobacter sp.* ADP1	0.65	0.18	0.39				0.46	0.05	0.60			
	E. coli K-12	0.82	0.24	0.21				0.66	0.09	0.36			
	F. novicida U112	0.65	0.20	0.48				0.55	0.08	0.66			
	M. genitalium G37	0.50	0.18	0.92				0.39	0.10	0.94			
	M. pulmonis UAB CTIP	0.59	0.12	0.77	0.64	0.18	0.55	0.50	0.06	0.84	0.51	0.08	0.68
90	*Acinetobacter sp.* ADP1	0.74	0.24	0.35				0.52	0.07	0.57			
	E. coli K-12	0.85	0.30	0.18				0.73	0.12	0.31			
	F. novicida U112	0.73	0.29	0.42				0.64	0.13	0.58			
	M. genitalium G37	0.60	0.24	0.91				0.52	0.13	0.94			
	M. pulmonis UAB CTIP	0.71	0.18	0.72	0.73	0.25	0.52	0.62	0.10	0.81	0.60	0.11	0.64
85	*Acinetobacter sp.* ADP1	0.77	0.28	0.33				0.54	0.08	0.53			
	E. coli K-12	0.87	0.34	0.16				0.76	0.14	0.29			
	F. novicida U112	0.76	0.34	0.40				0.67	0.16	0.55			
	M. genitalium G37	0.64	0.26	0.91				0.56	0.14	0.94			
	M. pulmonis UAB CTIP	0.74	0.21	0.70	0.76	0.29	0.50	0.67	0.12	0.78	0.64	0.13	0.62
80	*Acinetobacter sp.* ADP1	0.79	0.31	0.31				0.56	0.09	0.52			
	E. coli K-12	0.89	0.36	0.16				0.78	0.16	0.28			
	F. novicida U112	0.77	0.36	0.38				0.68	0.17	0.54			
	M. genitalium G37	0.66	0.32	0.89				0.59	0.18	0.93			
	M. pulmonis UAB CTIP	0.79	0.23	0.70	0.78	0.32	0.49	0.72	0.14	0.77	0.67	0.15	0.61

Continued on next page

Table A.2 – *Continued from previous page*

Cov.	Predicted species	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$
75	*Acinetobacter sp.* ADP1	0.80	0.33	0.30				0.57	0.10	0.50			
	E. coli K-12	0.89	0.38	0.15				0.78	0.16	0.27			
	F. novicida U112	0.78	0.39	0.37				0.70	0.18	0.54			
	M. genitalium G37	0.69	0.33	0.89				0.62	0.20	0.92			
	M. pulmonis UAB CTIP	0.80	0.24	0.69	0.79	0.33	0.48	0.73	0.15	0.76	0.68	0.16	0.60
70	*Acinetobacter sp.* ADP1	0.81	0.34	0.30				0.58	0.11	0.49			
	E. coli K-12	0.90	0.39	0.15				0.79	0.17	0.26			
	F. novicida U112	0.78	0.40	0.36				0.71	0.19	0.52			
	M. genitalium G37	0.70	0.36	0.88				0.64	0.23	0.92			
	M. pulmonis UAB CTIP	0.82	0.25	0.68	0.80	0.35	0.47	0.75	0.16	0.76	0.69	0.17	0.59
65	*Acinetobacter sp.* ADP1	0.82	0.36	0.29				0.60	0.12	0.48			
	E. coli K-12	0.91	0.41	0.15				0.81	0.18	0.26			
	F. novicida U112	0.79	0.42	0.35				0.71	0.20	0.52			
	M. genitalium G37	0.71	0.37	0.88				0.65	0.25	0.91			
	M. pulmonis UAB CTIP	0.83	0.26	0.67	0.81	0.36	0.47	0.77	0.16	0.76	0.71	0.18	0.58
60	*Acinetobacter sp.* ADP1	0.83	0.37	0.29				0.61	0.12	0.47			
	E. coli K-12	0.91	0.42	0.14				0.81	0.19	0.25			
	F. novicida U112	0.79	0.43	0.35				0.71	0.21	0.50			
	M. genitalium G37	0.72	0.40	0.88				0.65	0.26	0.91			
	M. pulmonis UAB CTIP	0.83	0.27	0.67	0.82	0.38	0.46	0.78	0.17	0.76	0.71	0.19	0.58
55	*Acinetobacter sp.* ADP1	0.84	0.38	0.28				0.61	0.13	0.46			
	E. coli K-12	0.92	0.43	0.14				0.82	0.20	0.24			
	F. novicida U112	0.80	0.45	0.34				0.72	0.22	0.49			
	M. genitalium G37	0.73	0.41	0.87				0.66	0.27	0.91			
	M. pulmonis UAB CTIP	0.84	0.28	0.66	0.82	0.39	0.46	0.79	0.17	0.75	0.72	0.20	0.57

Continued on next page

Table A.2 – *Continued from previous page*

Cov.	Predicted species	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$
50	*Acinetobacter sp.* ADP1	0.84	0.39	0.28				0.62	0.13	0.46			
	E. coli K-12	0.92	0.44	0.14				0.82	0.20	0.24			
	F. novicida U112	0.80	0.46	0.34				0.72	0.22	0.49			
	M. genitalium G37	0.73	0.41	0.87				0.66	0.28	0.90			
	M. pulmonis UAB CTIP	0.85	0.29	0.65	0.83	0.40	0.46	0.79	0.18	0.74	0.72	0.20	0.57
45	*Acinetobacter sp.* ADP1	0.85	0.40	0.27				0.62	0.13	0.45			
	E. coli K-12	0.92	0.45	0.14				0.83	0.21	0.23			
	F. novicida U112	0.81	0.47	0.34				0.73	0.23	0.48			
	M. genitalium G37	0.74	0.44	0.87				0.68	0.28	0.90			
	M. pulmonis UAB CTIP	0.85	0.30	0.65	0.83	0.41	0.45	0.79	0.18	0.74	0.73	0.21	0.56
40	*Acinetobacter sp.* ADP1	0.85	0.41	0.27				0.62	0.14	0.45			
	E. coli K-12	0.92	0.46	0.13				0.83	0.22	0.23			
	F. novicida U112	0.81	0.47	0.34				0.73	0.24	0.48			
	M. genitalium G37	0.74	0.44	0.87				0.69	0.29	0.90			
	M. pulmonis UAB CTIP	0.85	0.30	0.65	0.84	0.42	0.45	0.80	0.18	0.74	0.73	0.21	0.56
35	*Acinetobacter sp.* ADP1	0.85	0.41	0.27				0.63	0.14	0.44			
	E. coli K-12	0.92	0.47	0.13				0.83	0.22	0.22			
	F. novicida U112	0.81	0.48	0.33				0.73	0.24	0.47			
	M. genitalium G37	0.75	0.45	0.87				0.69	0.30	0.90			
	M. pulmonis UAB CTIP	0.85	0.31	0.64	0.84	0.43	0.45	0.80	0.19	0.74	0.74	0.22	0.56
30	*Acinetobacter sp.* ADP1	0.86	0.42	0.26				0.64	0.14	0.44			
	E. coli K-12	0.93	0.47	0.13				0.83	0.23	0.22			
	F. novicida U112	0.81	0.49	0.33				0.73	0.24	0.47			
	M. genitalium G37	0.75	0.46	0.86				0.70	0.31	0.90			
	M. pulmonis UAB CTIP	0.86	0.31	0.64	0.84	0.43	0.45	0.81	0.19	0.74	0.74	0.22	0.55

Continued on next page

Table A.2 – *Continued from previous page*

Cov.	Predicted species	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$
25	*Acinetobacter sp.* ADP1	0.86	0.43	0.26				0.64	0.15	0.43			
	E. coli K-12	0.93	0.48	0.13				0.83	0.23	0.21			
	F. novicida U112	0.81	0.49	0.33				0.73	0.25	0.46			
	M. genitalium G37	0.75	0.47	0.86				0.70	0.31	0.90			
	M. pulmonis UAB CTIP	0.87	0.32	0.64	0.84	0.44	0.44	0.82	0.19	0.74	0.74	0.23	0.55
20	*Acinetobacter sp.* ADP1	0.86	0.44	0.26				0.64	0.15	0.43			
	E. coli K-12	0.93	0.49	0.13				0.83	0.24	0.21			
	F. novicida U112	0.81	0.50	0.32				0.73	0.26	0.46			
	M. genitalium G37	0.76	0.47	0.86				0.70	0.32	0.90			
	M. pulmonis UAB CTIP	0.87	0.32	0.64	0.84	0.44	0.44	0.82	0.20	0.73	0.74	0.23	0.54
15	*Acinetobacter sp.* ADP1	0.87	0.44	0.26				0.64	0.15	0.43			
	E. coli K-12	0.93	0.49	0.13				0.83	0.25	0.21			
	F. novicida U112	0.81	0.50	0.32				0.73	0.26	0.45			
	M. genitalium G37	0.76	0.48	0.86				0.70	0.32	0.90			
	M. pulmonis UAB CTIP	0.87	0.33	0.64	0.85	0.45	0.44	0.82	0.20	0.73	0.75	0.24	0.54
10	*Acinetobacter sp.* ADP1	0.87	0.44	0.26				0.64	0.16	0.42			
	E. coli K-12	0.93	0.49	0.13				0.83	0.25	0.20			
	F. novicida U112	0.81	0.50	0.32				0.73	0.26	0.45			
	M. genitalium G37	0.76	0.48	0.86				0.71	0.32	0.90			
	M. pulmonis UAB CTIP	0.87	0.33	0.64	0.85	0.45	0.44	0.82	0.20	0.73	0.75	0.24	0.54
5	*Acinetobacter sp.* ADP1	0.87	0.44	0.26				0.65	0.16	0.42			
	E. coli K-12	0.93	0.49	0.13				0.83	0.25	0.20			
	F. novicida U112	0.81	0.50	0.32				0.73	0.26	0.45			
	M. genitalium G37	0.76	0.48	0.86				0.71	0.32	0.90			

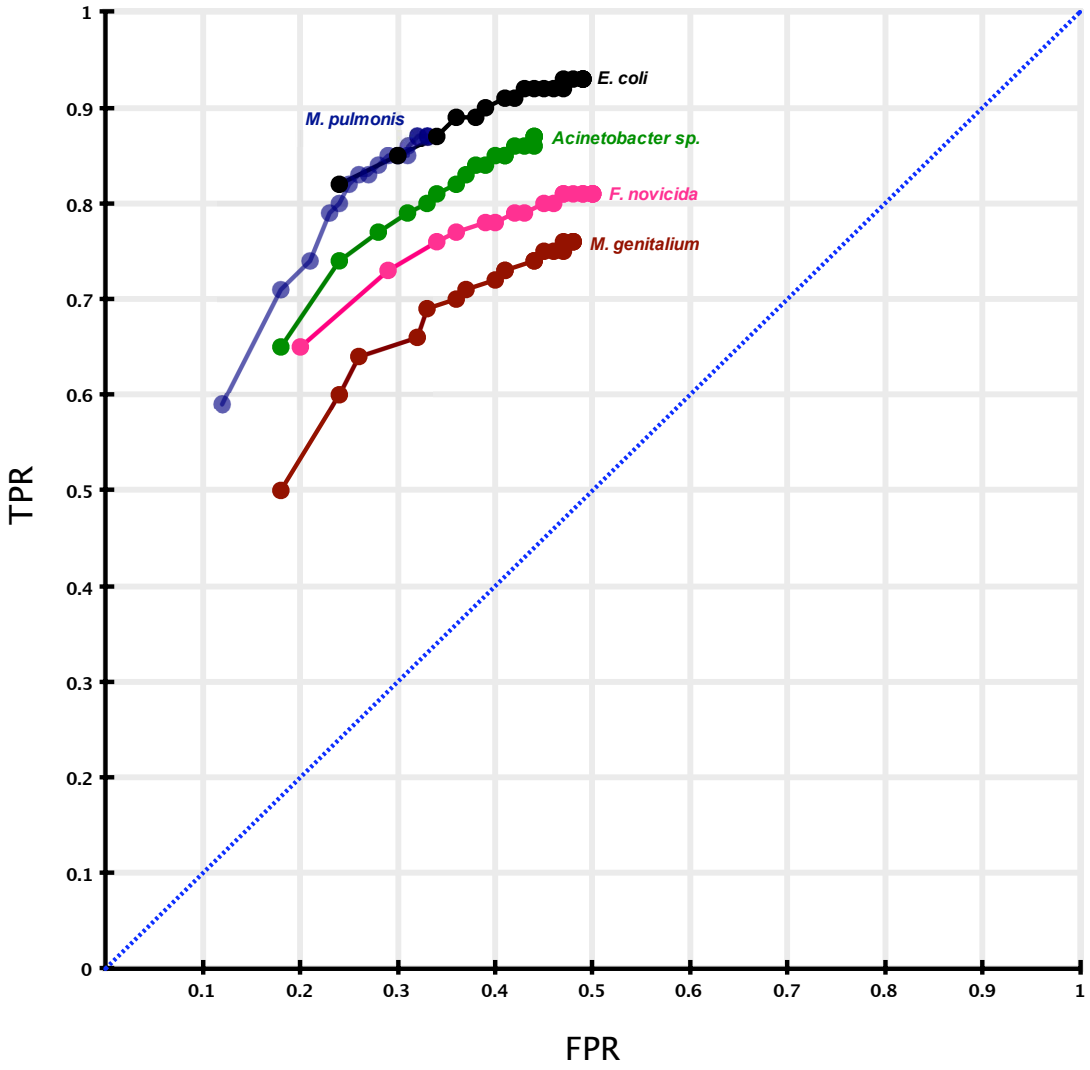
Continued on next page

Table A.2 – *Continued from previous page*

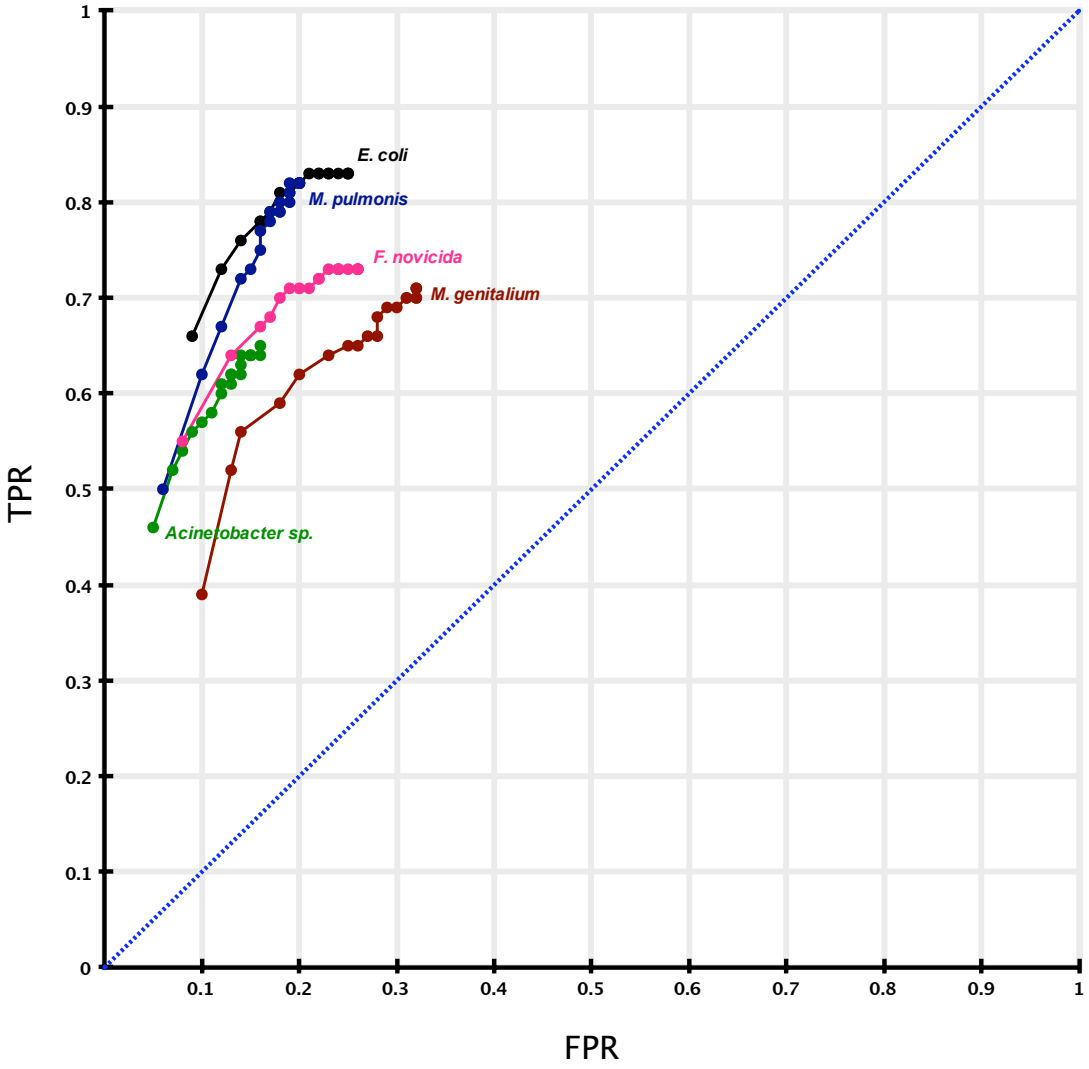
Cov.	Predicted species	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$	TPR	FPR	PPV	$\overline{TPR}$	$\overline{FPR}$	$\overline{PPV}$
	M. pulmonis UAB CTIP	0.87	0.33	0.64	0.85	0.45	0.44	0.82	0.20	0.73	0.75	0.24	0.54

Figure A.1: Homology benchmark details. Performance of predicting essential proteins in five species using homology to two databases of known essential proteins.

(a) Performance using the Database of essential genes (DEG).

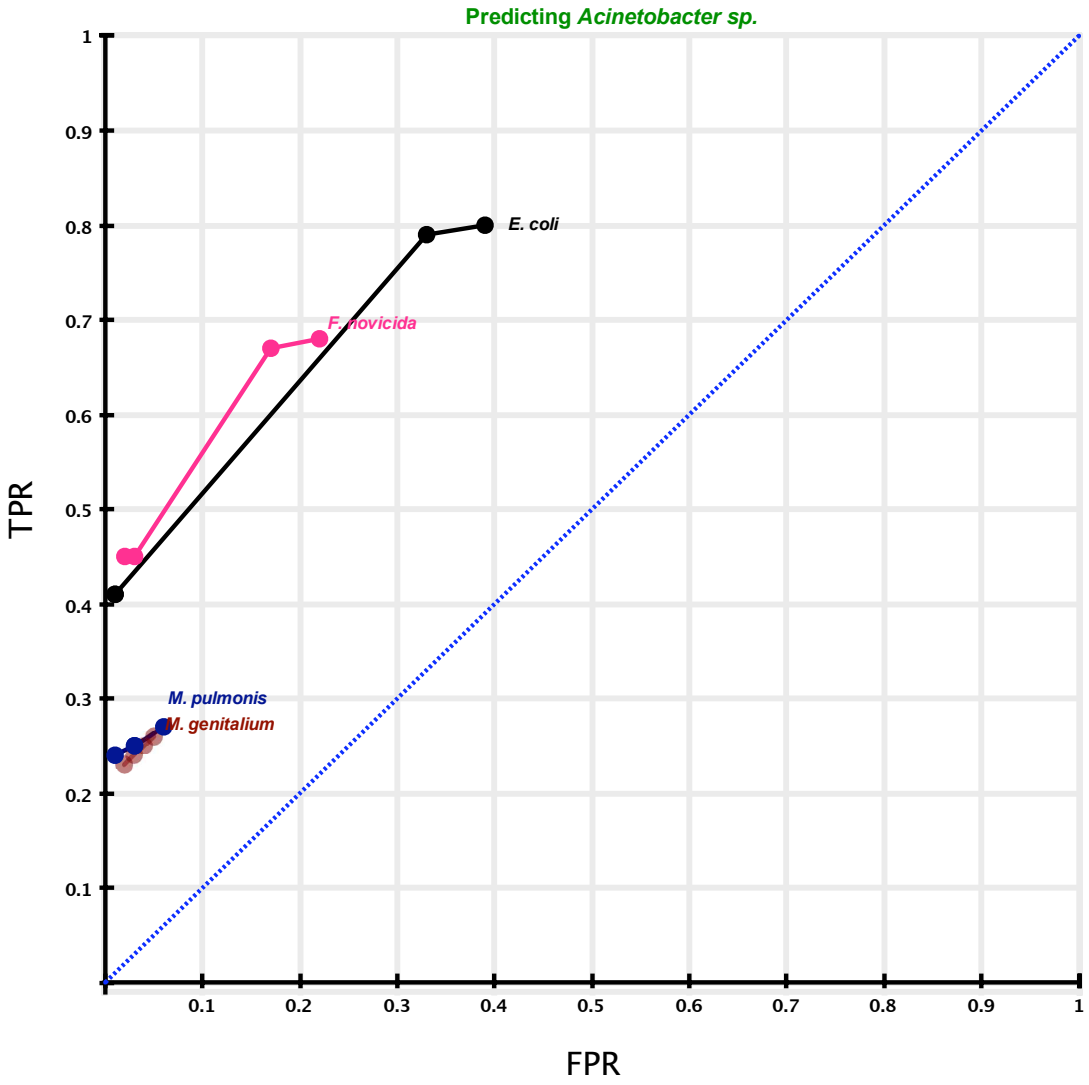


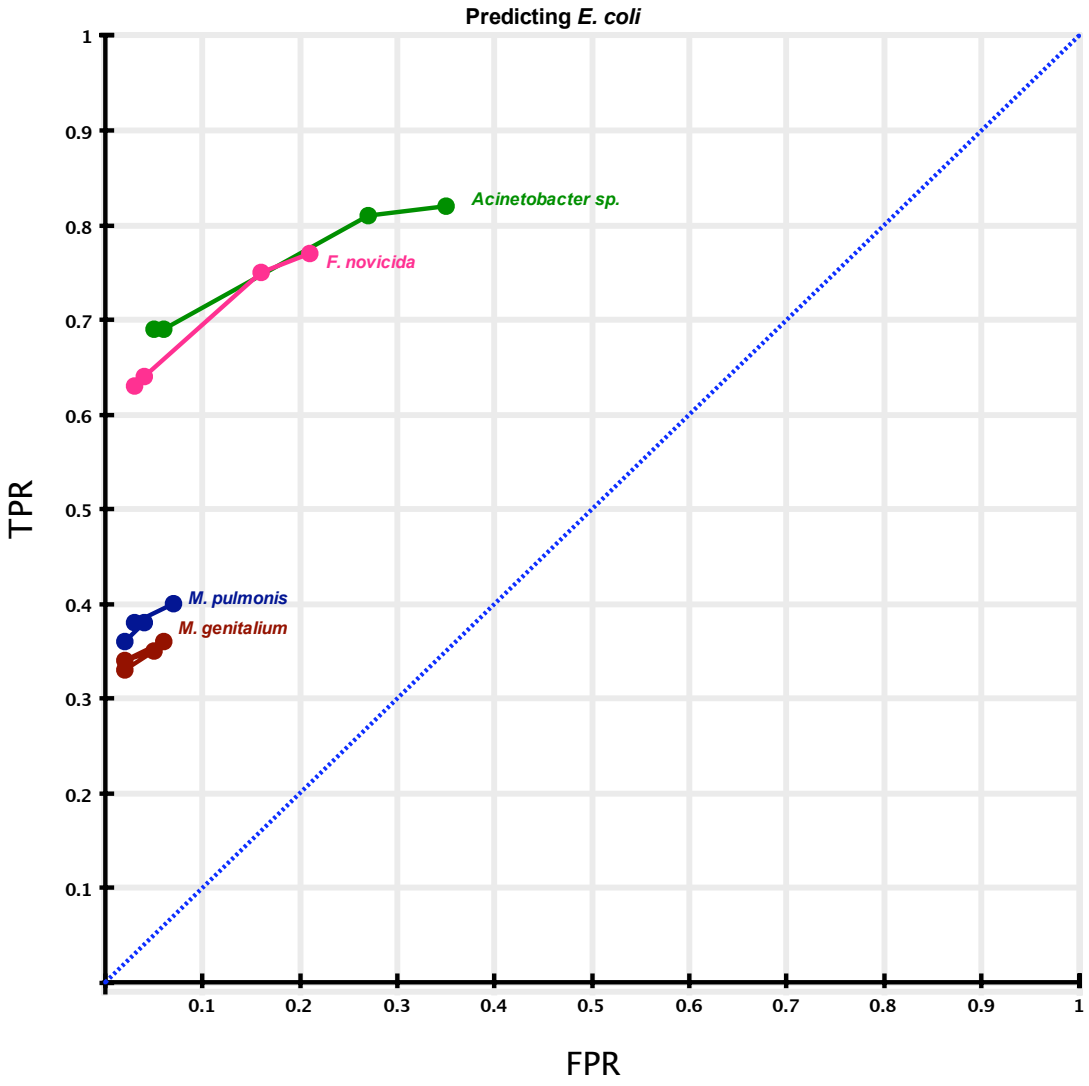
(b) Performance using four essential proteomes.

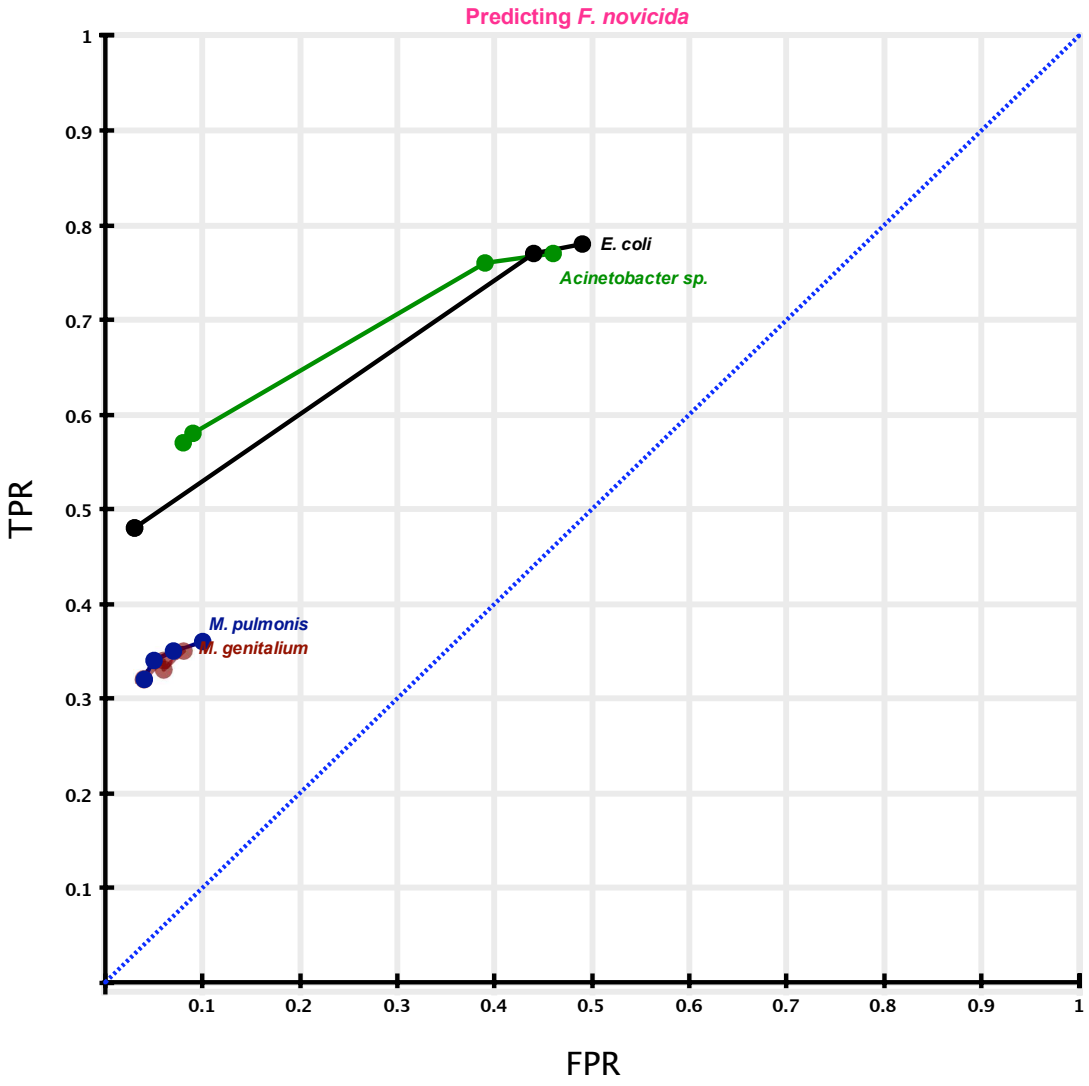


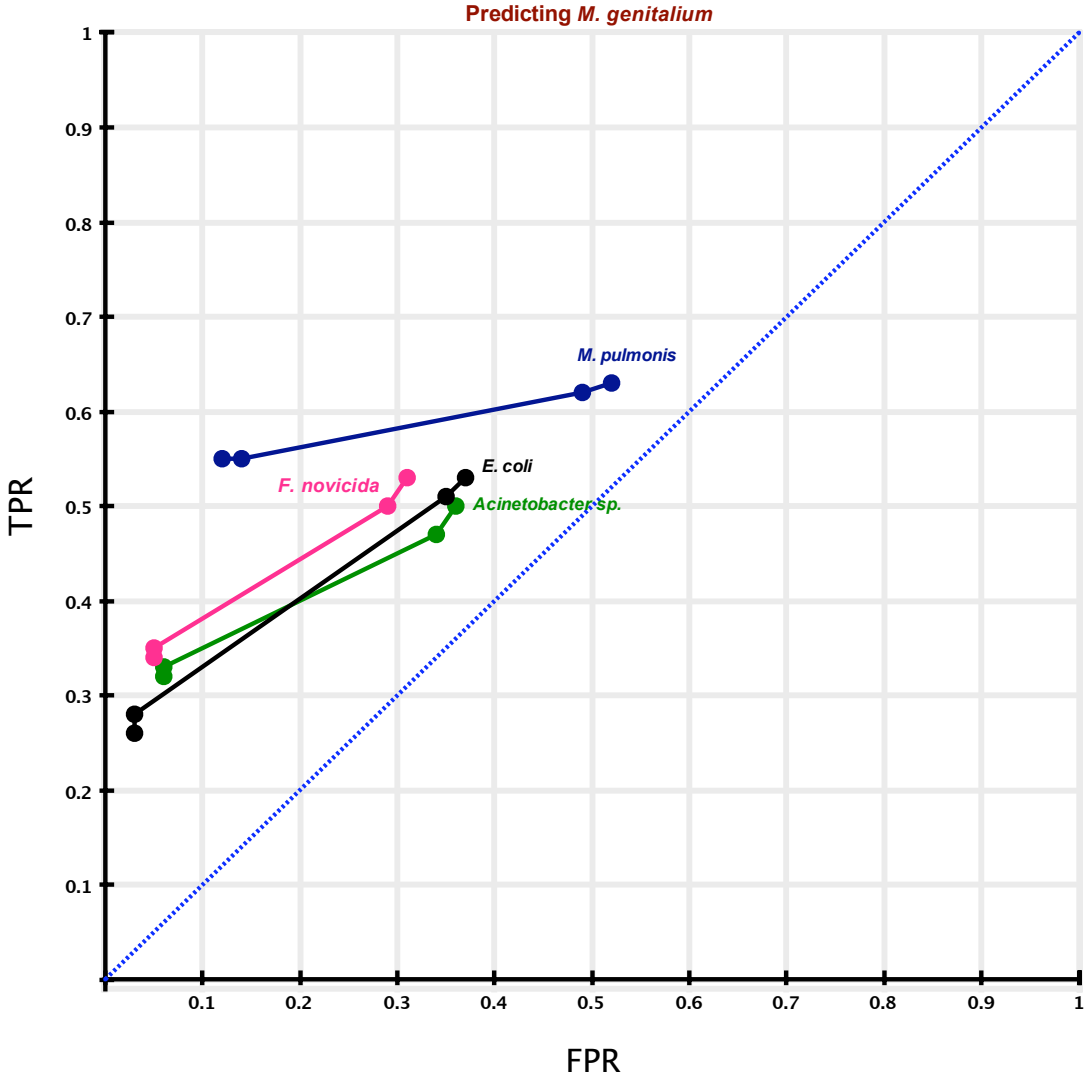
A.3 Orthology inferrence

Figure A.2: Orthology benchmark details. Performance of predicting essential proteins in five species using orthology models. The four models of essentiallity shown here are m1, m2, m3 and m4. Models are connected in numerical order starting with m1, which always produces the largest FPR.




Predicting E. coli
TPR
FPR
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Acinetobacter sp.
F. novicida
M. pulmonis
M. genitalium


Predicting F. novicida
TPR
FPR
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
E. coli
Acinetobacter sp.
M. pulmonis
M. genitalium


Predicting M. genitalium
TPR
FPR
M. pulmonis
E. coli
F. novicida
Acinetobacter sp.
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

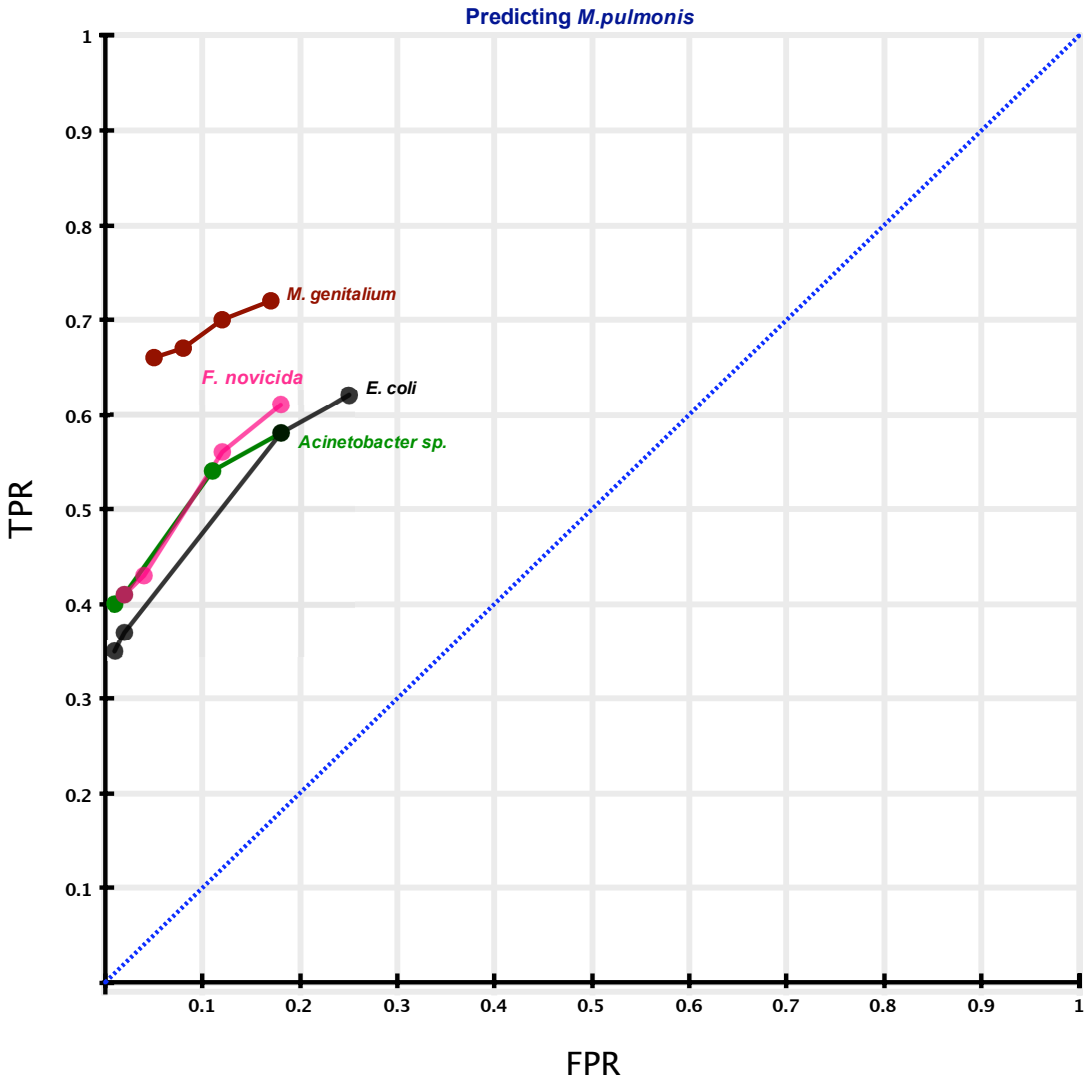

Predicting M.pulmonis
TPR
FPR
M. genitalium
F. novicida
E. coli
Acinetobacter sp.
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure A.3: Orthology benchmark details. Performance of predicting essential proteins in five species using orthology models. The four models of essentiality shown here are m5(1), m5(2), m5(3) and m5(4). Models are connected in numerical order starting with m5(1), which always produces the largest FPR.

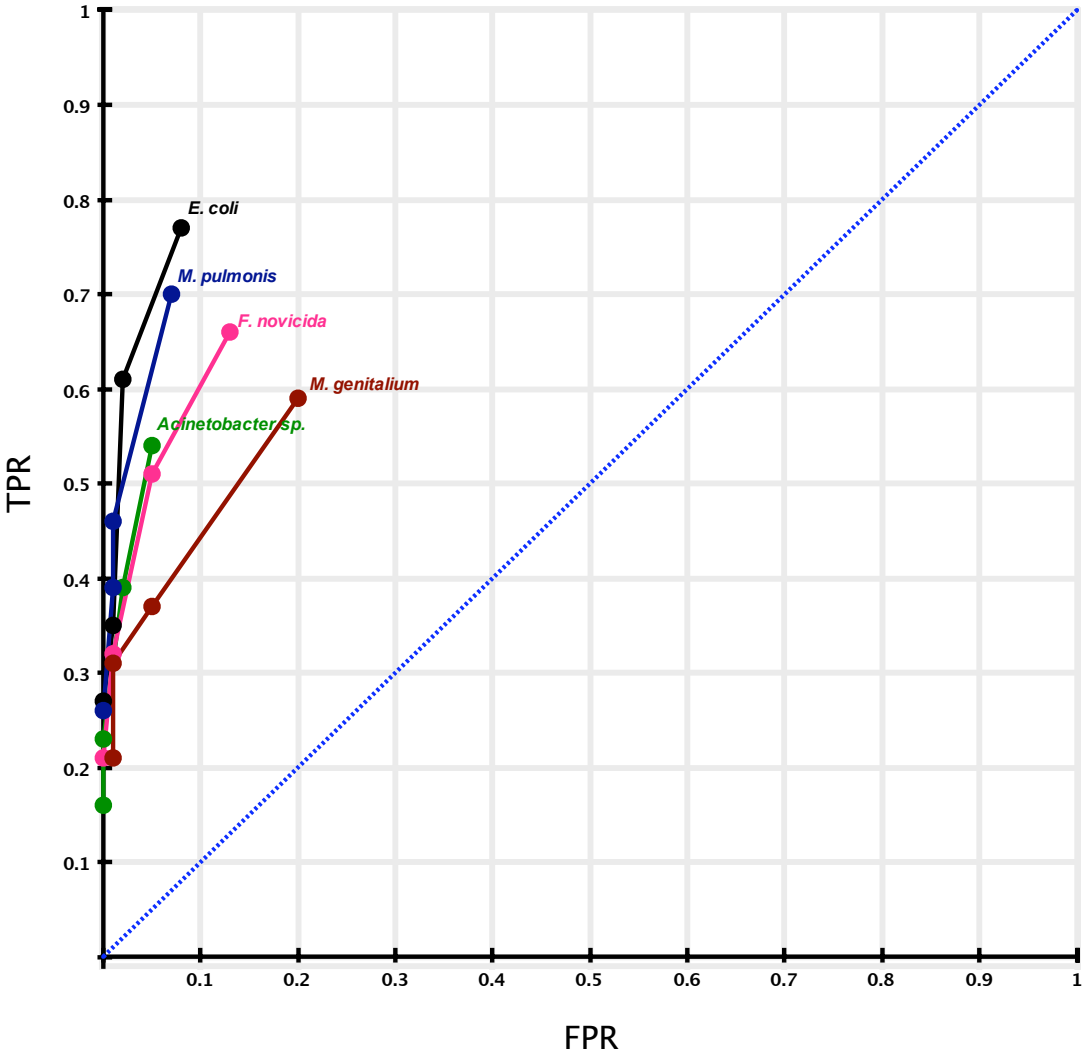


Table A.3: The proteins predicted essential in any of the five benchmark species by the most specific model - m5(4).

Function	Protein description	Gene name(s)
tRNA metabolism	alanyl-tRNA synthetase	alaS
	arginyl-tRNA synthetase	argS
	aspartyl-tRNA synthetase	aspS
	cysteinyl-tRNA synthetase	cysS
	glutamyl-tRNA synthetase	gltX
	histidyl-tRNA synthetase	hisS
	isoleucyl-tRNA synthetase	ileS
	leucyl-tRNA synthetase	leuS
	tRNA(Ile)-lysidine synthetase	mesJ,tilS
	methionyl-tRNA synthetase	metG,metS
	seryl-tRNA synthetase	serS
	threonyl-tRNA synthetase	thrS
	tryptophanyl-tRNA synthetase	trpS
	tyrosyl-tRNA synthetase	tyrS
	valyl-tRNA synthetase	valS
	phenylalanyl-tRNA synthetase subunit alpha	pheS
	phenylalanyl-tRNA synthetase subunit beta	pheT
	tRNA (guanine-N(1)-)-methyltransferase	trmD
	methionyl-tRNA formyltransferase	arnA,fmt
	peptidyl-tRNA hydrolase	pth
DNA metabolism	DNA primase	dnaE,dnaG,polC-2
	DNA topoisomerase I	topA
	DNA polymerase III subunits gamma and tau	dnaX
	DNA-directed RNA polymerase subunit alpha	rpoA
	DNA-directed RNA polymerase subunit beta	rpoB
	DNA-directed RNA polymerase subunit beta	rpoC
	DNA gyrase subunit A	gyrA
	DNA gyrase subunit B	gyrB,parE
	replicative DNA helicase	dnaB
	NAD-dependent DNA ligase LigA	ligA
	dihydrofolate reductase	dhfR,folA
Protein synthesis	peptide chain release factor 1	prfA
	preprotein translocase subunit SecA	secA
	preprotein translocase subunit SecY	secY
	translation initiation factor IF-1	infA

Continued on next page

Table A.3 – *Continued from previous page*

Function	Protein description	Gene name(s)
	translation initiation factor IF-2	infB
	translation initiation factor IF-3	infC
	elongation factor G	fus,fusA
	elongation factor Ts	tsf
	ribosome biogenesis GTP-binding protein YsxC	engB,yihA
	ribosome recycling factor	frr
	50S ribosomal protein L1	rplA
	50S ribosomal protein L10	rplJ
	50S ribosomal protein L11	rplK
	50S ribosomal protein L13	rplM
	50S ribosomal protein L14	rplN
	50S ribosomal protein L15	rplO
	50S ribosomal protein L16	rplP
	50S ribosomal protein L17	rplQ
	50S ribosomal protein L18	rplR
	50S ribosomal protein L19	rplS
	50S ribosomal protein L2	rplB
	50S ribosomal protein L20	rplT
	50S ribosomal protein L21	rplU
	50S ribosomal protein L22	rplV
	50S ribosomal protein L23	rplW
	50S ribosomal protein L24	rplX
	50S ribosomal protein L27	rpmA
	50S ribosomal protein L3	rplC
	50S ribosomal protein L34	rpmH
	50S ribosomal protein L4	rplD
	50S ribosomal protein L5	rplE
	50S ribosomal protein L6	rplF
	50S ribosomal protein L7/L12	rplL
	30S ribosomal protein S10	rpsJ
	30S ribosomal protein S11	rpS11,rpsK
	30S ribosomal protein S12	rpsL
	30S ribosomal protein S13	rpsM
	30S ribosomal protein S14	rpsN
	30S ribosomal protein S15	rpsO
	30S ribosomal protein S16	rpsP
	30S ribosomal protein S17	rpsQ

Continued on next page

Table A.3 – *Continued from previous page*

Function	Protein description	Gene name(s)
	30S ribosomal protein S18	rpsR
	30S ribosomal protein S19	rpsS
	30S ribosomal protein S2	rpsB
	30S ribosomal protein S3	rpsC
	30S ribosomal protein S4	rpsD
	30S ribosomal protein S5	rpsE
	30S ribosomal protein S7	rpS7,rpsG
	30S ribosomal protein S8	rpsH
	30S ribosomal protein S9	rpsI
Transcription	RNA polymerase sigma factor RpoD	rpoD
	transcription elongation factor NusA	nusA
	transcription antitermination protein NusG	nusG
Energy metabolism	triosephosphate isomerase	tpiA
	F0F1 ATP synthase subunit alpha	atpA
	F0F1 ATP synthase subunit beta	atpD
	F0F1 ATP synthase subunit gamma	atpG
	adenylate kinase	adk
Cell division	cell division protein FtsH	ftsH,hflB
	cell division protein FtsY	ftsY
	cell division protein FtsZ	ftsZ
	chromosomal replication initiation protein	dnaA
Transportation	signal recognition particle protein	ffh
	macrolide transporter ATP-binding /permease protein	lolD,macB,ybbA
Other	NAD synthetase	nadE
	cytidylate kinase	cmk
	peptide deformylase	def
	phosphopyruvate hydratase	eno
	glyceraldehyde-3-phosphate dehydrogenase	gap,gapA
	serine hydroxymethyltransferase	glyA
	guanylate kinase	gmk
	S-adenosylmethionine synthetase	metK,metX
	GTPase ObgE	obgE
	phosphoglycerate kinase	pgk
	phosphatidylglycerophosphate synthetase	pgsA
	inorganic pyrophosphatase	ppa
	ribose-phosphate pyrophosphokinase	prs,prsA

Continued on next page

Table A.3 – *Continued from previous page*

Function	Protein description	Gene name(s)
	uridylate kinase	pyrH
	putative DNA-binding/iron metalloprotein/AP endonuclease	gcp,ygjD
	phosphoglyceromutase	gpmI,pgm
	GTP-binding protein EngA	der,engA
	thymidylate kinase	tmk

A.4 Kinetoplastids

		BF^{D3} ∪ BF^{D6}		**BF^{D6}**		**BF^{D3}**		**BF^{D3} ∩ BF^{D6}**	
kinetoplastid	**Psize**	m3	**m4**	m3	**m4**	m3	**m4**	m3	**m4**
T. cruzi Esmeraldo-Like	10342	2587	**2219**	2289	**1944**	1613	**1319**	1314	**1043**
T. brucei gambiense	9668	3072	**2832**	2704	**2474**	1886	**1715**	1515	**1356**
T. brucei Lister strain 427	8529	3255	**2837**	2868	**2480**	2032	**1698**	1643	**1341**
L. infantum	8033	2527	**2342**	2240	**2065**	1573	**1427**	1285	**1149**
L. major	8045	2510	**2349**	2222	**2073**	1559	**1431**	1270	**1154**
L. braziliensis	7809	2446	**2266**	2168	**1999**	1530	**1382**	1251	**1114**

Table A.4: Numbers of predicted essential proteins in 6 kinetoplastids. (where **Psize** = proteome size; BF^{D3}/BF^{D6} = bloodstream form after 3/6 days respectivly). The essential proteins of *T. bruci* strain TREU 927. were used to infer essentiality using models m3 and m4 (as described in Chapter 2.2.7). The essential predictions were based on proteins shown to be essential in either BF^{D6}, BF^{D3}, both forms or either form.

kinetoplastid	Psize	% of proteome
T. cruzi Esmeraldo-Like	10342	61 (6323)
T. brucei gambiense	9668	80 (7739)
T. brucei Lister strain 427	8529	98 (8345)
L. infantum	8033	70 (5653)
L. major	8045	70 (5657)
L. braziliensis	7809	70 (5499)

Table A.5: Percentage of kinetoplastid proteme with co-ortholog relationships with *T. bruci* strain TREU 927. (Psize = proteome size; Protein count in brackets).

A.5 Matrix database

Table A.6: The database tables for protein family sequence information and screen compounds activity information. The datatypes for each column are shown. Each database tables primary key (where available) is shown (**pk**). The reference table(s) that post foreign keys are shown in **fk table**. The **notes** section described the column data with example where appropriate.

matrix_alignment - describes a protein family multiple sequence alignment (MSA).				
column name	**type**	**pk**	**fk table**	**notes**
Alignment_id	String	Y		Unique name given to the protein family alignment e.g. "HumanAndPDBKinases".
Accession	String	Y	matrix_gene_to_uniprot	UniProt accession.
Seq_res	String			The amino acid residue at this MSA position e.g. "-" (alignment gap).
Aln_pos	Number	Y		The sequential position in the MSA.
Seq_pos	Number			The sequential position in the UniProt sequence.
Domain_id	Number	Y		The unique identifier for this domain, important where the UniProt contains repeats of the same domain family.
matrix_sites - describes a binding site of a protein family in terms of sequence positions.				
column name	**type**	**pk**	**fk table**	**notes**
Site_id	String	Y		unique name given to the binding site e.g. "conserved_atp_site".
Alignment_id	String	Y	matrix_ alignment	unique name given to the protein family alignment e.g. "HumanAndPDBKinases".
Accession	String	Y	matrix_ alignment	UniProt accession of a family member used to describe the binding site.
Seq_pos	Number	Y	matrix_ alignment	UniProt sequence position.
matrix_gene_to_uniprot - maps gene names and synonyms to UniProtKB accessions.				
column name	**type**	**pk**	**fk table**	**notes**
Taxon	Number	Y	newt_entry	NEWT taxon identifier for the protein screened.

Continued on next page

Table A.6 – *Continued from previous page*

Accession	String	Y		UniProtKB protein accession.
Synonym	String	Y		the gene name or synonym for the UniProt protein.
matrix_activities - describes a set of protein-compound screening data.				
column name	**type**	**pk**	**fk table**	**notes**
Taxon	Number	Y	newt_entry	NEWT taxon identifier for the protein screened.
External_data_source	String	Y		source of activities data e.g. "Abbott".
External_compound_id	String	Y		unique identifier given for the compound.
Activity_units	String			measurement of affinity e.g "pK_i"
Activity_value	Number			affinity of the compound to the protein.
Activity_qualifier	String			used to qualify the affinity e.g. "<" (affinity lower than activity_value).
External_protein_id	String	Y	matrix_gene_to_uniprot .synonym	identifier given for the protein (expected to be a standard gene name).
matrix_compounds - describes the compounds of a assay.				
column name	**type**	**pk**	**fk table**	**notes**
External_data_source	String	Y	matrix_ activities	source of activities data e.g. "Abbott".
Cannonical_smiles	String			string representation of the compound structure.
External_compound_id	String	Y	matrix_ activities	unique identifier for the compound given by source.
matrix_aaindex - describes a the AAindex resource. a database of amino acid descriptors.				
column name	**type**	**pk**	**fk table**	**notes**
Accession	String	Y		External AAindex accession. e.g. "ARGP820101".
Seq_res	String	Y		The amino acid residue described e.g. "Y" (Tyrosine).
Original_value	Number			The index value for the amino acid e.g. "1.88".
Normal_value	Number			The index value for the amino acid when scale normalized between -1 and 1.

Continued on next page

Table A.6 – *Continued from previous page*

Positive_shifted	Number			The index value for the amino acid when scale normalized between 1 and 10.

A.5.0.2 Matrix database usage examples

Consult a compound screening experiment to find (a) how many compounds have been screened against some proteins of interest. Then add an acitivity cutoff to find (b) only the potent compounds.

```
SELECT gu.accession                    uniprot,
       ma.external_protein_id          gene,
       COUNT(ma.external_compound_id) compounds
  FROM matrix_activities      ma
  JOIN matrix_gene_to_uniprot gu
    ON ma.external_protein_id  = gu.synonym
 WHERE gu.accession           IN ('P00519','O00311','O75582')
   -- use the Abbott screening data
   AND ma.external_data_source = 'abbott_kinase'
   -- we know Abbott use pKi, so set an "active" cutoff
   AND ma.activity_value       >= 7       -- remove for (a)
   -- ignore screens where activity not resolved
   AND (ma.activity_qualifier != '<'
    OR  ma.activity_qualifier IS NULL)   -- remove for (a)
 GROUP BY gu.accession, ma.external_protein_id;
```

(a) shows some genes have been screened with many more compounds than others:

```
UNIPROT |GENE    |COMPOUNDS
---------------------------
O75582  |RPS6KA5 | 938
O00311  |CDC7    |1912
P00519  |ABL1    |2065
```

(b) shows more of the compounds screened are potent against CDC7:

```
UNIPROT |GENE    |COMPOUNDS
---------------------------
O75582  |RPS6KA5 | 51
```

```
O00311   |CDC7     |806
P00519   |ABL1     |262
```

Find the positions in the MSA which reflect a ligand binding site.

```
SELECT ma.aln_pos
  FROM matrix_sites ms
  JOIN matrix_alignment ma
    ON ms.alignment_id = ma.alignment_id
 WHERE ms.accession    = ma.accession
   AND ms.seq_pos      = ma.seq_pos
   -- a small ligand binding site observed from
   -- a binding site profile clustering run
   AND ms.site_id      = 'UnCentCorAverage_16'
   -- the MSA the clustering used
   AND ms.alignment_id = 'HumanAndCredo1';
```

Which produces 7 MSA postions:

```
ALN_POS
-------
569
670
671
672
674
1458
1459
1460
```

Use this small binding site definition to find the equivalent residues from another protein, and use an amino acid index to describe these residues:

```
SELECT ma.domain_id,
       ma.seq_pos  ,
       ma.seq_res  ,
       mi.original_value
  FROM matrix_alignment ma
  JOIN matrix_aaindex   mi
    ON mi.seq_res = ma.seq_res
 WHERE ma.aln_pos in (569,670,671,672,674,1458,1459,1460)
   AND ma.alignment_id = 'HumanAndCredo1'
```

```
    -- the UniProt protein we are describing
    AND ma.accession        = 'O00141'
    -- an index of hydrophobicity
    AND mi.accession = 'ARGP820101'
 ORDER BY domain_id, ma.aln_pos;
```

produces a numerical value describing the hyrophobicity of the binding site residues. In this case the protein target had only 1 kinase domain. The sequence postions of the binding site on the original UniProtKB protein sequence are also shown:

```
DOMAIN_ID |SEQ_POS |SEQ_RES |ORIGINAL_VALUE
-----------------------------------------
1         |189     |Q       |0
1         |191     |E       |0.47
1         |192     |R       |0.6
1         |193     |C       |1.07
1         |195     |L       |1.53
1         |292     |Y       |1.88
1         |293     |G       |0.07
1         |294     |L       |1.53
```

A.5.1 Binding site analysis

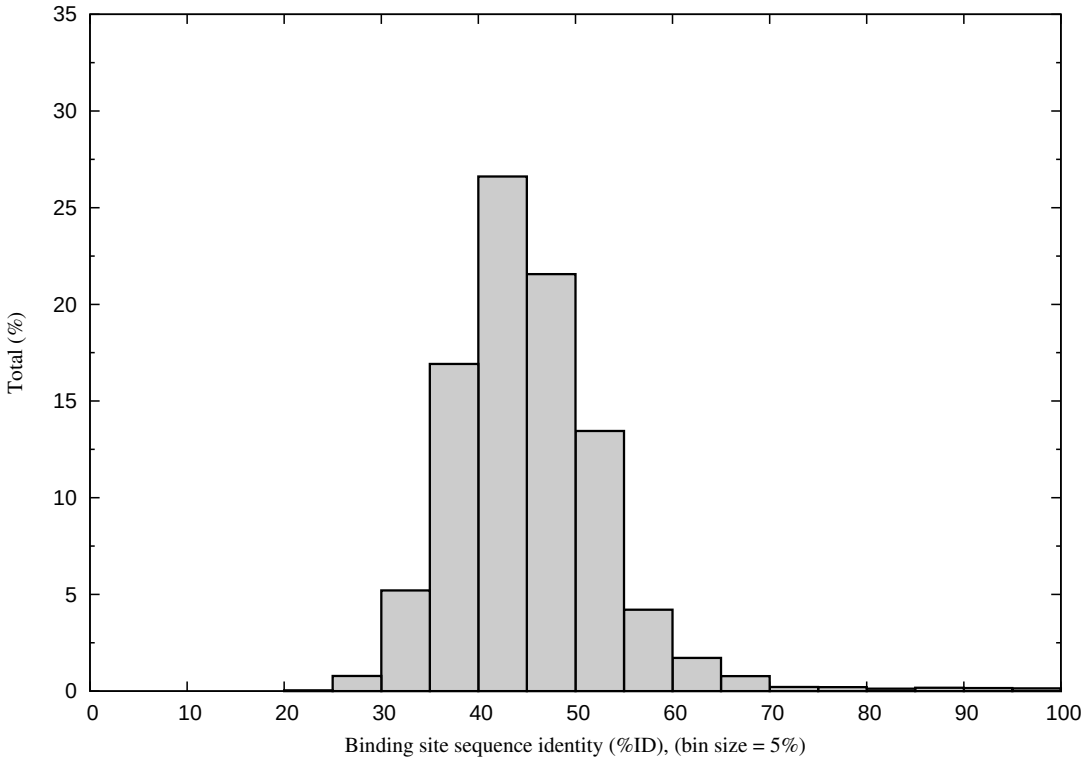
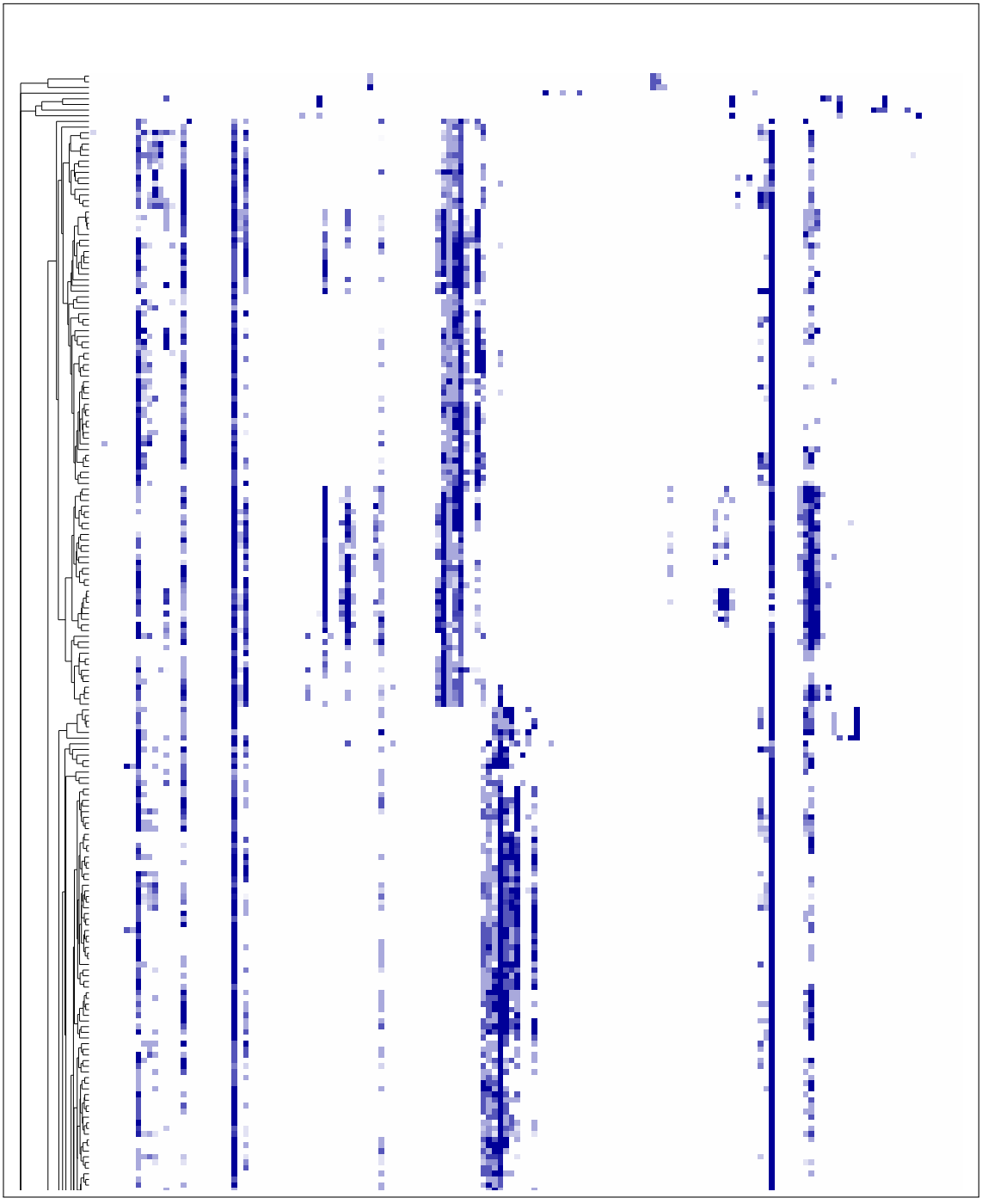
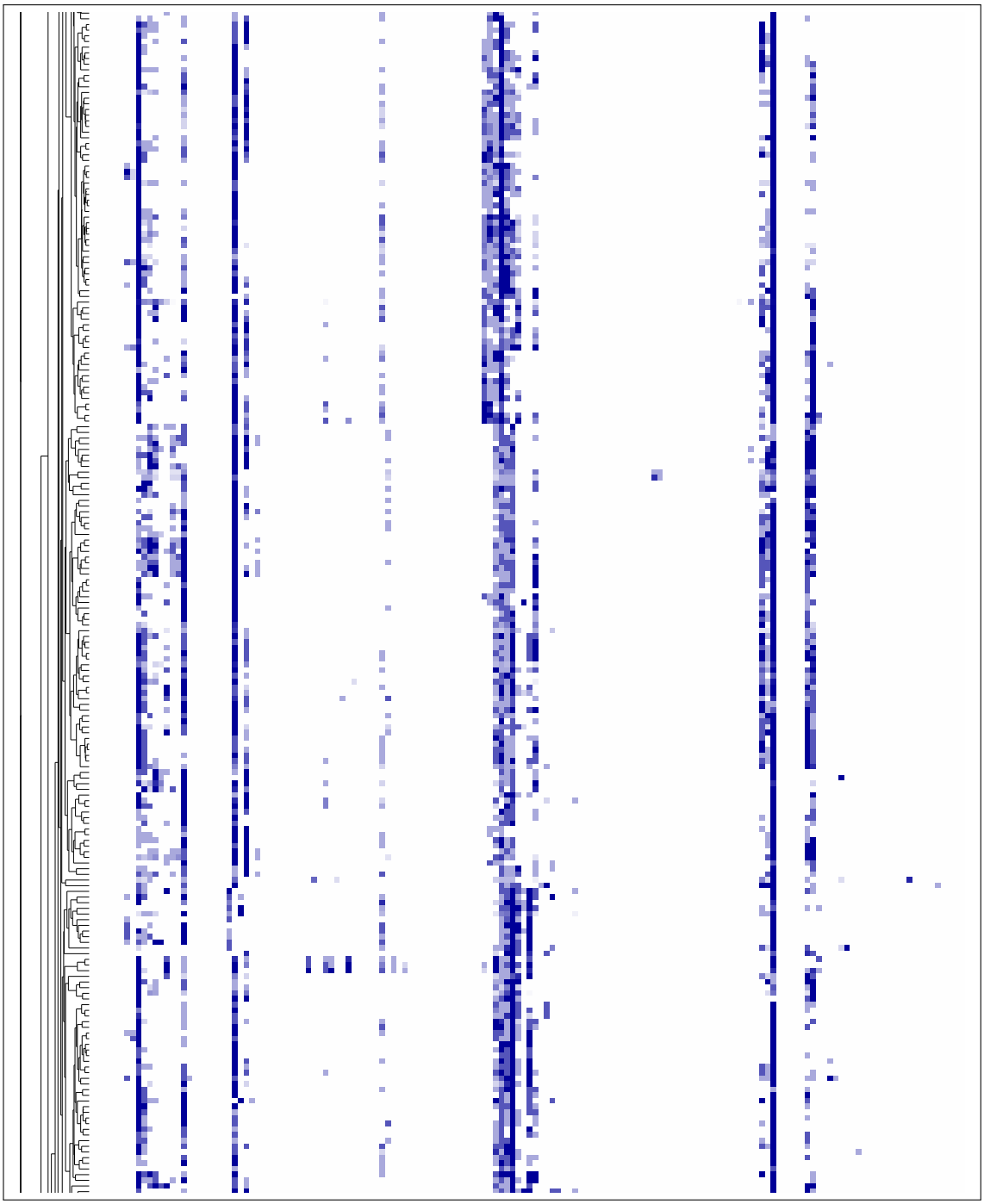
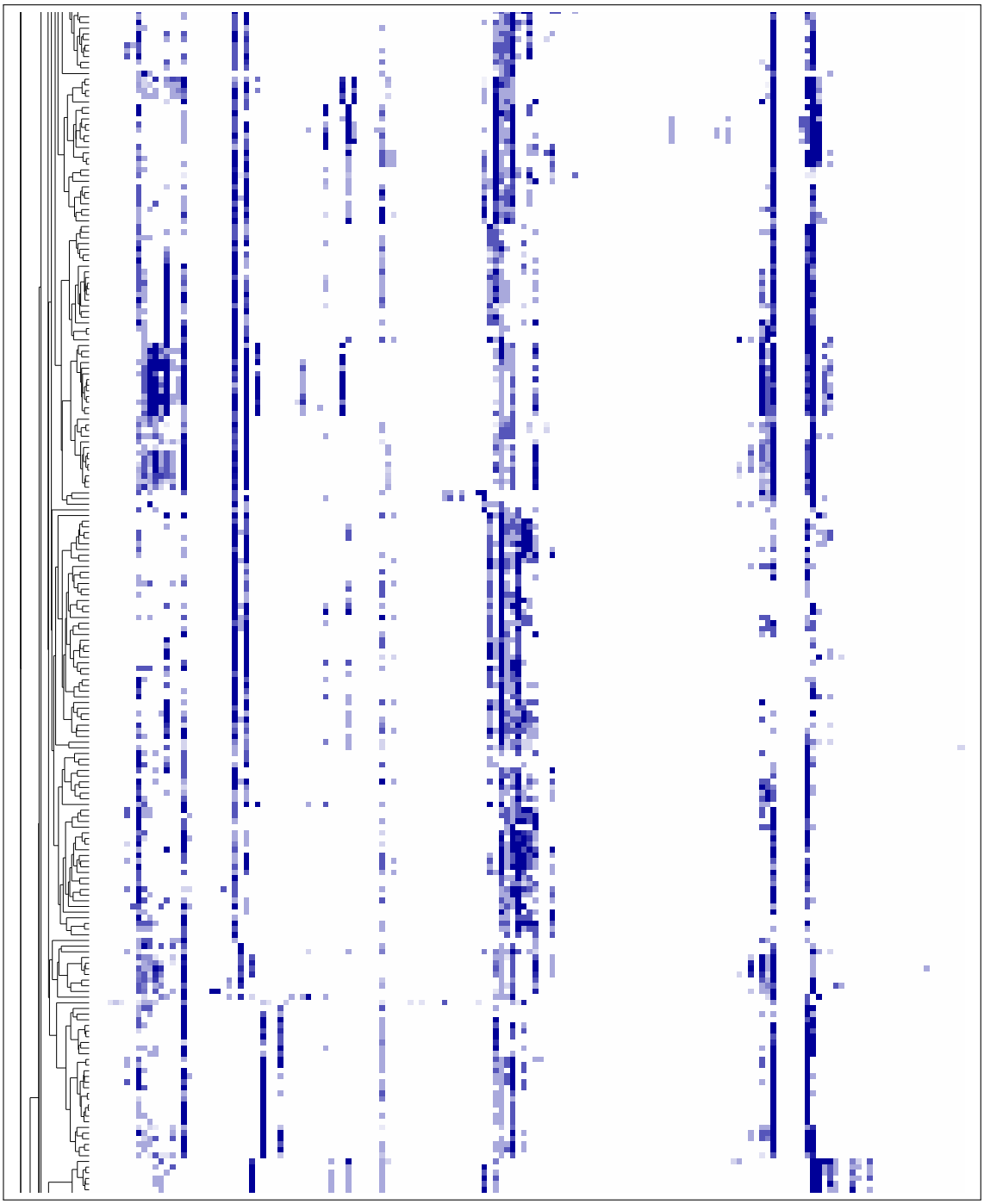


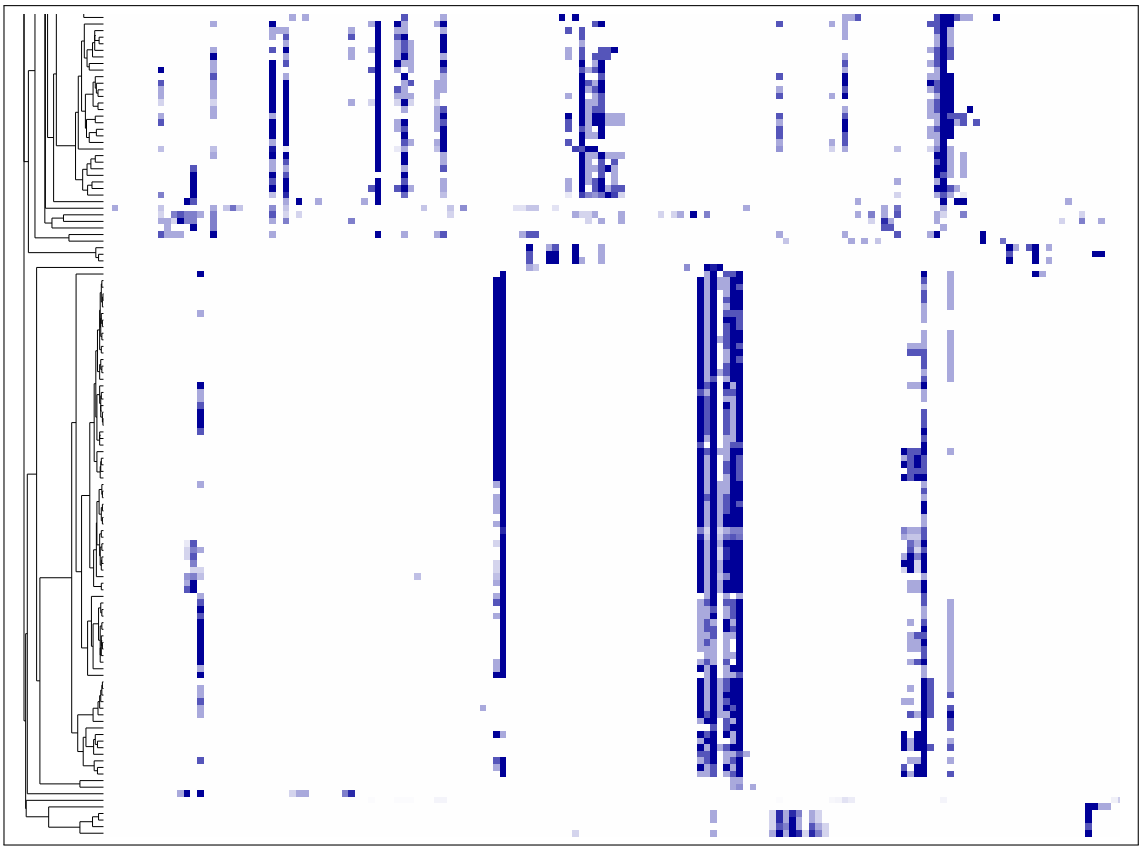
Figure A.4: The distribution of binding site sequence identity in the Protein kinases. The sequence identity (%ID) of the kinase pairs was calculated over the 48 residues in the Loose ATP binding site.

Figure A.5: Hierarchical clustering of the PLIP highlight multiple modes of ligand binding. Modes largely overlap with the natural ATP binding, but distinct allosteric sites can be observed. Clustering performed using Cluster 3.0 software (de Hoon *et al.*, 2004). Figure generated using TreeView (Saldanha, 2004).









A.5.2 CREDO database

In this work, the major function of the CREDO database was to a) find the binding sites of pdb ligands and b) to map those binding sites onto UniProtKB sequences (so the binding sites information could be transferred to all protein family members via a multiple sequence alignment). The following query finds all interesting ligands of a kinase pdb chain, and maps their binding residues onto a PLIP:

```
SELECT DISTINCT
       rmp.uniprot_accession   ,
       lig.name    ligand_name ,
       rmp.res_num uniprot_res ,
       rmp.one_letter_code  olc,
       count(distinct con.hetatm_id) contacts
  FROM credo.structures str
  JOIN credo.chains     chn
    ON str.id         = chn.structure_id
  JOIN credo.residues   res
    ON chn.id         = res.chain_id
  JOIN credo.atoms      atm
    ON res.id         = atm.residue_id
  JOIN credo.contacts   con
    ON atm.id         = con.atom_id
  JOIN credo.hetatms    het
    ON con.hetatm_id = het.id
  JOIN credo.ligands    lig
    ON het.ligand_id = lig.id
  JOIN credo.residuemap rmp
    ON res.id         = rmp.residue_id
  -- ignore is_proximal types
 WHERE con.is_proximal   = 0
  -- ignore very small ligands of <= 8 heavy atoms
   AND lig.num_hvy_atoms > 8
  -- this is out query PDB entry
   AND str.pdb      = '3EQH'
   AND chn.pdb_chain_id = 'A'
```

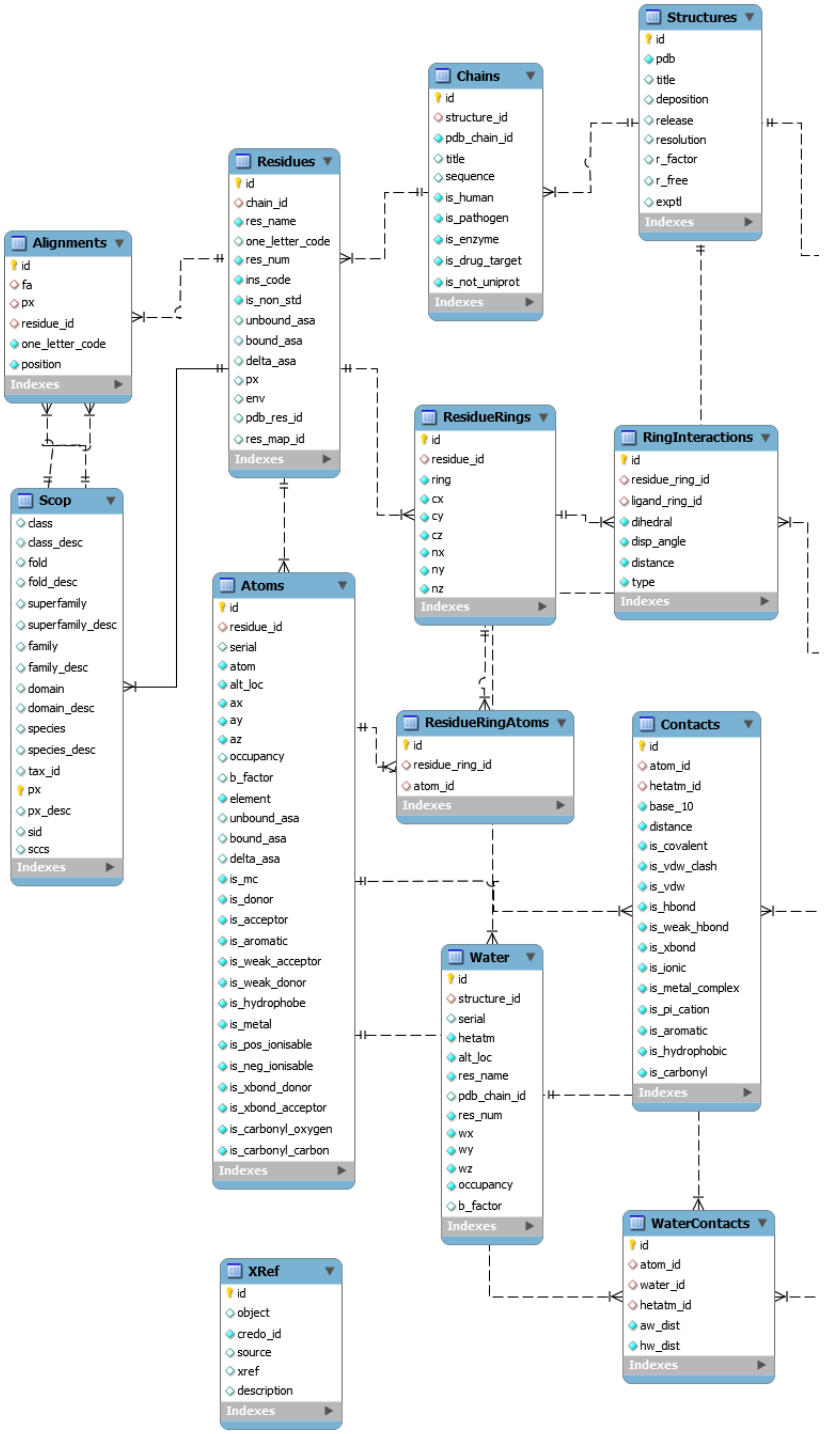
```
GROUP BY
   rmp.uniprot,
   lig.name,
   rmp.res_num,
   rmp.one_letter_code
ORDER BY ligand_name,uni_res_num;
```

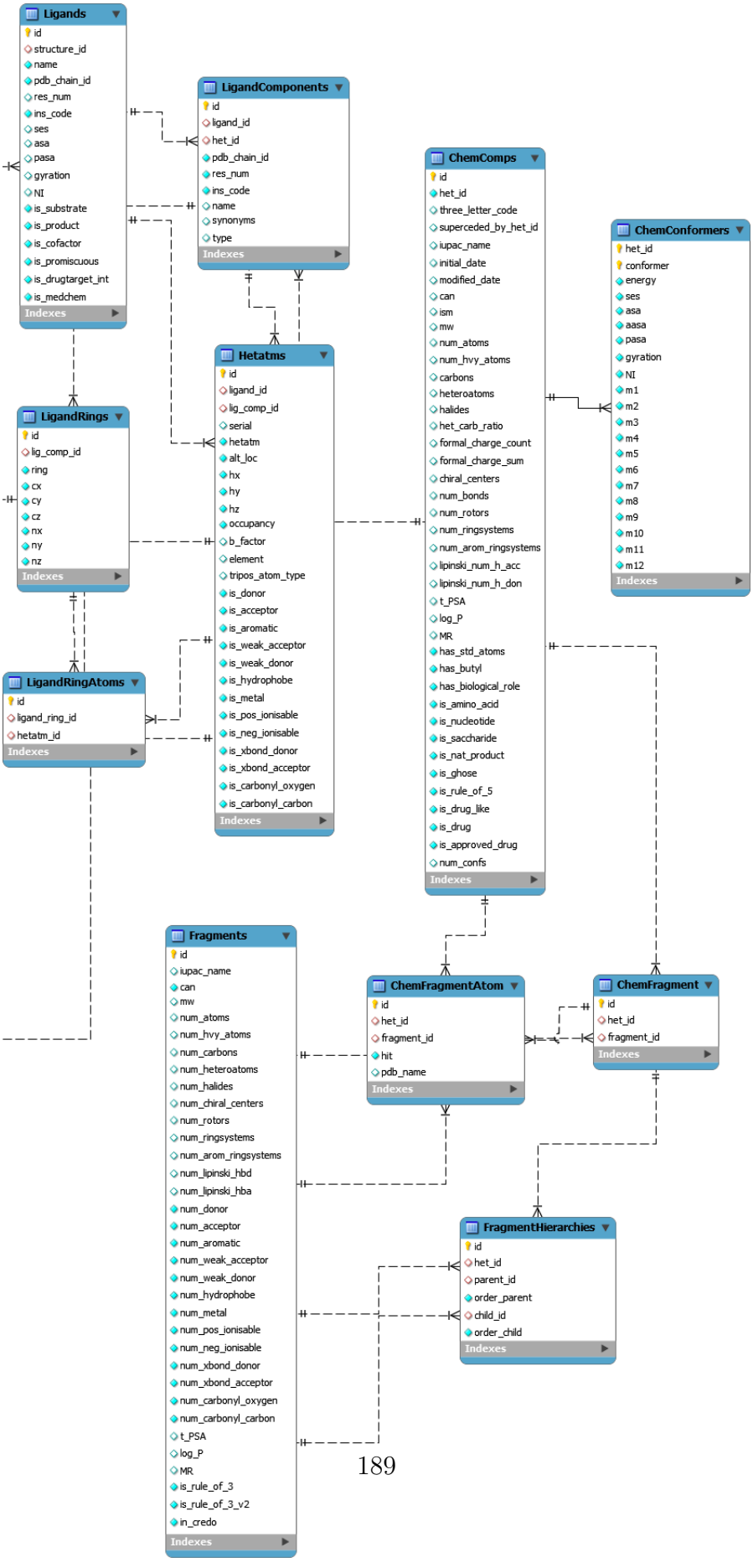
Finds two ligands for this pdb entry (both also shown in Figute 5.4), each of these ligands sites could be mapped across the Kinome using the matrix database described above (see A.5.0.2):

```
ACCESSION  |LIGAND_NAME |UNIPROT_RES |OLC |CONTACTS
-------------------------------------------------
Q02750     |5BM         |97          |K   |1
Q02750     |5BM         |99          |I   |1
Q02750     |5BM         |118         |L   |2
Q02750     |5BM         |127         |V   |1
Q02750     |5BM         |141         |I   |3
Q02750     |5BM         |143         |M   |3
Q02750     |5BM         |188         |H   |1
Q02750     |5BM         |189         |R   |1
Q02750     |5BM         |190         |D   |2
Q02750     |5BM         |207         |C   |2
Q02750     |5BM         |208         |D   |6
Q02750     |5BM         |209         |F   |5
Q02750     |5BM         |210         |G   |2
Q02750     |5BM         |211         |V   |1
Q02750     |5BM         |212         |S   |1
Q02750     |5BM         |216         |I   |4
-------------------------------------------------
Q02750     |ADP         |74          |L   |2
Q02750     |ADP         |75          |G   |1
Q02750     |ADP         |77          |G   |1
Q02750     |ADP         |80          |G   |1
Q02750     |ADP         |82          |V   |4
Q02750     |ADP         |95          |A   |2
Q02750     |ADP         |97          |K   |2
Q02750     |ADP         |144         |E   |1
Q02750     |ADP         |145         |H   |1
```

```
Q02750      |ADP          |146          |M    |2
Q02750      |ADP          |150          |S    |2
Q02750      |ADP          |153          |Q    |1
Q02750      |ADP          |192          |K    |1
Q02750      |ADP          |194          |S    |3
Q02750      |ADP          |195          |N    |1
Q02750      |ADP          |197          |L    |4
Q02750      |ADP          |208          |D    |1
```

CREDO Schema Diagram on two pages, (created by Adrian Schreyer). The Oracle implementation mirrors this MySQL version where possible. More information about the data content of CREDO can be found at http://marid.bioc.cam.ac.uk/credo.


Structures
id
pdb
title
deposition
release
resolution
r_factor
r_free
exptl
Indexes
Chains
id
structure_id
pdb_chain_id
title
sequence
is_human
is_pathogen
is_enzyme
is_drug_target
is_not_uniprot
Indexes
Residues
id
chain_id
res_name
one_letter_code
res_num
ins_code
is_non_std
unbound_asa
bound_asa
delta_asa
px
env
pdb_res_id
res_map_id
Indexes
Alignments
id
fa
px
residue_id
one_letter_code
position
Indexes
Scop
class
class_desc
fold
fold_desc
superfamily
superfamily_desc
family
family_desc
domain
domain_desc
species
species_desc
tax_id
px
px_desc
sid
sccs
Indexes
ResidueRings
id
residue_id
ring
cx
cy
cz
nx
ny
nz
Indexes
RingInteractions
id
residue_ring_id
ligand_ring_id
dihedral
disp_angle
distance
type
Indexes
Atoms
id
residue_id
serial
atom
alt_loc
ax
ay
az
occupancy
b_factor
element
unbound_asa
bound_asa
delta_asa
is_mc
is_donor
is_acceptor
is_aromatic
is_weak_acceptor
is_weak_donor
is_hydrophobe
is_metal
is_pos_ionisable
is_neg_ionisable
is_xbond_donor
is_xbond_acceptor
is_carbonyl_oxygen
is_carbonyl_carbon
Indexes
ResidueRingAtoms
id
residue_ring_id
atom_id
Indexes
Contacts
id
atom_id
hetatm_id
base_10
distance
is_covalent
is_vdw_clash
is_vdw
is_hbond
is_weak_hbond
is_xbond
is_ionic
is_metal_complex
is_pi_cation
is_aromatic
is_hydrophobic
is_carbonyl
Indexes
Water
id
structure_id
serial
hetatm
alt_loc
res_name
pdb_chain_id
res_num
wx
wy
wz
occupancy
b_factor
Indexes
WaterContacts
id
atom_id
water_id
hetatm_id
aw_dist
hw_dist
Indexes
XRef
id
object
credo_id
source
xref
description
Indexes


Ligands
id
structure_id
name
pdb_chain_id
res_num
ins_code
ses
asa
pasa
gyration
NI
is_substrate
is_product
is_cofactor
is_promiscuous
is_drugtarget_int
is_medchem
Indexes
LigandComponents
id
ligand_id
het_id
pdb_chain_id
res_num
ins_code
name
synonyms
type
Indexes
ChemComps
id
het_id
three_letter_code
superceded_by_het_id
iupac_name
initial_date
modified_date
can
ism
mw
num_atoms
num_hvy_atoms
carbons
heteroatoms
halides
het_carb_ratio
formal_charge_count
formal_charge_sum
chiral_centers
num_bonds
num_rotors
num_ringsystems
num_arom_ringsystems
lipinski_num_h_acc
lipinski_num_h_don
t_PSA
log_P
MR
has_std_atoms
has_butyl
has_biological_role
is_amino_acid
is_nucleotide
is_saccharide
is_nat_product
is_ghose
is_rule_of_5
is_drug_like
is_drug
is_approved_drug
num_confs
Indexes
ChemConformers
het_id
conformer
energy
ses
asa
aasa
pasa
gyration
NI
m1
m2
m3
m4
m5
m6
m7
m8
m9
m10
m11
m12
Indexes
Hetatms
id
ligand_id
lig_comp_id
serial
hetatm
alt_loc
hx
hy
hz
occupancy
b_factor
element
tripos_atom_type
is_donor
is_acceptor
is_aromatic
is_weak_acceptor
is_weak_donor
is_hydrophobe
is_metal
is_pos_ionisable
is_neg_ionisable
is_xbond_donor
is_xbond_acceptor
is_carbonyl_oxygen
is_carbonyl_carbon
Indexes
LigandRings
id
lig_comp_id
ring
cx
cy
cz
nx
ny
nz
Indexes
LigandRingAtoms
id
ligand_ring_id
hetatm_id
Indexes
Fragments
id
iupac_name
can
mw
num_atoms
num_hvy_atoms
num_carbons
num_heteroatoms
num_halides
num_chiral_centers
num_rotors
num_ringsystems
num_arom_ringsystems
num_lipinski_hbd
num_lipinski_hba
num_donor
num_acceptor
num_aromatic
num_weak_acceptor
num_weak_donor
num_hydrophobe
num_metal
num_pos_ionisable
num_neg_ionisable
num_xbond_donor
num_xbond_acceptor
num_carbonyl_oxygen
num_carbonyl_carbon
t_PSA
log_P
MR
is_rule_of_3
is_rule_of_3_v2
in_credo
Indexes
ChemFragmentAtom
id
het_id
fragment_id
hit
pdb_name
Indexes
ChemFragment
id
het_id
fragment_id
Indexes
FragmentHierarchies
id
het_id
parent_id
order_parent
child_id
order_child
Indexes

References

Aguero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., Carmona, S., Carruthers, I.M., Chan, A.W.E., Chen, F., Crowther, G.J., Doyle, M.A., Hertz-Fowler, C., Hopkins, A.L., McAllister, G., Nwaka, S., Overington, J.P., Pain, A., Paolini, G.V., Pieper, U., Ralph, S.A., Riechers, A., Roos, D.S., Sali, A., Shanmugam, D., Suzuki, T., Van Voorhis, W.C. & Verlinde, C.L.M.J. (2008). Genomic-scale prioritization of drug targets: the TDR Targets database. *Nature Reviews Drug Discovery*, **7**, 900–907. 6, 23, 50, 81, 104, 145

Akerley, B.J., Rubin, E.J., Camilli, A., Lampe, D.J., Robertson, H.M. & Mekalanos, J.J. (1998). Systematic identification of essential genes by in vitro mariner mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8927–8932. 21

Al-Lazikani, B., Gaulton, A., Paolini, G., Lanfear, J., Overington, J. & Hopkins, A. (2007). The Molecular Basis of Predicting Druggability. In G. Wess & S. Schreiber, eds., *Chemical Biology*, Wiley. 147

Alsford, S., Turner, D.J., Obado, S.O., Sanchez-Flores, A., Glover, L., Berriman, M., Hertz-Fowler, C. & Horn, D. (2011). High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome research*, **21**, 915–924. 21, 69

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410. 24

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S. & Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish development. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12792–12797. 77

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B.P., Carrington, M., Depledge, D.P., Fischer, S., Gajria, B., Gao, X., Gardner, M.J., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Logan, F.J., Miller, J.A., Mitra, S., Myler, P.J., Nayak, V., Pennington, C., Phan, I., Pinney, D.F., Ramasamy, G., Rogers, M.B., Roos, D.S., Ross, C., Sivam, D., Smith, D.F., Srinivasamoorthy, G., Stoeckert, C.J., Subramanian, S., Thibodeau, R., Tivey, A., Treatman, C., Velarde, G. & Wang, H. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research*, **38**, D457–D462. 69

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, **2**, msb4100050–E1–msb4100050–E11. 10, 32

Balaji, S., Sujatha, S., Kumar, S.C. & Srinivasan, N. (2001). PALIa database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Research*, **29**, 61–65. 108

Balcht, A. & Smith, R. (1994). *Pseudomonas Aeruginosa: Infections and Treatment.*. No. 12 in Infectious Disease and Therapy, Marcel Dekker. 64

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651. 74, 77

Becker, M., Aitcheson, N., Byles, E., Wickstead, B., Louis, E. & Rudenko, G. (2004). Isolation of the repertoire of VSG expression site con-

taining telomeres of Trypanosoma brucei 427 using transformation-associated recombination in yeast. *Genome research*, **14**, 2319–2329. 68

Bellis, L.J., Akhtar, R., AlLazikani, B., Atkinson, F., Bento, A.P., Chambers, J., Davies, M., Gaulton, A., Hersey, A., Ikeda, K., Krüger, F.A., Light, Y., McGlinchey, S., Santos, R., Stauch, B. & Overington, J.P. (2011). Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society Transactions*, **39**, 1365–1370. 5

Berglund, A.C.C., Sjölund, E., Ostlund, G. & Sonnhammer, E.L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic acids research*, **36**. 33

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242. 85

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Böhme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P.J., Moule, S., Martin, D.M., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabbinowitsch, E., Rajandream, M.A., Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M., Tait, A., Tivey, A.R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S.,

Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J.D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E. & El-Sayed, N.M. (2005). The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416–422. 68

Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B., Andrade, L.F., Ashton, P.D., Aslett, M.A., Bartholomeu, D.C., Blandin, G., Caffrey, C.R., Coghlan, A., Coulson, R., Day, T.A., Delcher, A., DeMarco, R., Djikeng, A., Eyre, T., Gamble, J.A., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D.A., Lacerda, D., Macedo, C.D., McVeigh, P., Ning, Z., Oliveira, G., Overington, J.P., Parkhill, J., Pertea, M., Pierce, R.J., Protasio, A.V., Quail, M.A., Rajandream, M.A.A., Rogers, J., Sajid, M., Salzberg, S.L., Stanke, M., Tivey, A.R., White, O., Williams, D.L., Wortman, J., Wu, W., Zamanian, M., Zerlotini, A., Fraser-Liggett, C.M., Barrell, B.G. & El-Sayed, N.M. (2009). The genome of the blood fluke Schistosoma mansoni. *Nature*, **460**, 352–358. 6, 73, 74, 75, 81

Bianchi, V., Gherardini, P.F., Citterich, M.H. & Ausiello, G. (2012). Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. *BMC Bioinformatics*, **13**, S17+. 9

Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S. & Hopkins, A.L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, **4**, 90–98. 5

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Paro, R., Perrimon, N. & Heidelberg Fly Array Consortium (2004). Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science (New York, N.Y.)*, **303**, 832–835. 77

Brötz-Oesterhelt, H. & Sass, P. (2010). Postgenomic strategies in antibacterial drug discovery. *Future Microbiology*, **5**, 1553–1579. 5

Brun, R., Blum, J., Chappuis, F. & Burri, C. (2010). Human African trypanosomiasis. *The Lancet*, **375**, 148–159. 69

Bundkirchen, A., Brixius, K., Bölck, B., Nguyen, Q. & Schwinger, R.H. (2003). 1-adrenoceptor selectivity of nebivolol and bisoprolol. A comparison of [3H]CGP 12.177 and [125I]iodocyanopindolol binding studies. *European Journal of Pharmacology*, **460**, 19–26. 105

Caffrey, C.R., Rohwer, A., Oellien, F., Marhöfer, R.J., Braschi, S., Oliveira, G., McKerrow, J.H. & Selzer, P.M. (2009). A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni. *PloS one*, **4**, e4413+. 6, 23

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421+. 34, 82

Case, K. (2013). OmniGraffle: for Mac. xiii, xviii, 28, 114

Chakrabarti, P. & Bhattacharyya, R. (2007). Geometry of nonbonded interactions involving planar groups in proteins. *Progress in biophysics and molecular biology*, **95**, 83–137. 111

Chan, M. (2007). Reaching the people left behind: A neglected success. In *Prince Mahidol Award Conference*. 2

Chan, P.F., Macarron, R., Payne, D.J., Zalacain, M. & Holmes, D.J. (2002). Novel antibacterials: a genomics approach to drug discovery. *Current drug targets. Infectious disorders*, **2**, 291–308. 102

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S.E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–192+. 87

Chapman, T.M. & Perry, C.M. (2012). Cefepime: A Review of its Use in the Management of Hospitalized Patients with Pneumonia. *American Journal of Respiratory Medicine*, **2**, 75–107. 19

Chen, F., Mackey, A.J., Vermunt, J.K. & Roos, D.S. (2007). Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLoS ONE*, **2**, e383+. xii, 26, 33, 35

Chen, W.H.H., Minguez, P., Lercher, M.J. & Bork, P. (2012). OGEE: an online gene essentiality database. *Nucleic acids research*, **40**, D901–D906. 27

Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H. & Shapiro, L. (2011). The essential genome of a bacterium. *Molecular systems biology*, **7**. 61, 145

Codd, E.F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, **13**, 377–387. 25

Curry, S. & Brown, R. (2003). The target product profile as a planning tool in drug discovery research. *Business Briefing: PharmaTech*, 67–71+. 6

Davies, J. & Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews*, **74**, 417–433. 2

de Berardinis, V., Vallenet, D., Castelli, V., Besnard, M., Pinet, A., Cruaud, C., Samair, S., Lechaplais, C., Gyapay, G., Richez, C., Durot, M., Kreimeyer, A., Le Fevre, F., Schachter, V., Pezo, V., Doring, V., Scarpelli, C., Medigue, C., Cohen, G.N., Marliere, P., Salanoubat, M. & Weissenbach, J. (2008). A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Molecular Systems Biology*, **4**. 30, 32

de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. (2004). Open source clustering software. *Bioinformatics (Oxford, England)*, **20**, 1453–1454. xxii, 124, 181

Delano, W.L. (2006). MacPyMOL: PyMOL enhanced for Mac OS X. xvi, xviii, xix, 84, 120, 125, 126

D'Elia, M.A., Pereira, M.P. & Brown, E.D. (2009). Are essential genes really essential? *Trends in Microbiology*, **17**, 433–438. 10

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A.A., Hassett, D.J. & Lu, L.J. (2011). Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Research*, **39**, 795–807. 62

Dengler, U., Siddiqui, A.S. & Barton, G.J. (2001). Protein structural domains: analysis of the 3Dee domains database. *Proteins*, **42**, 332–344. 83

Denome, S.A., Elf, P.K., Henderson, T.A., Nelson, D.E. & Young, K.D. (1999). Escherichia coli Mutants Lacking All Possible Combinations of Eight Penicillin Binding Proteins: Viability, Characteristics, and Implications for Peptidoglycan Synthesis. *Journal of Bacteriology*, **181**, 3981–3993. 10, 12, 19

Doenhoff, M.J., Hagan, P., Cioli, D., Southgate, V., Pica-Mattoccia, L., Botros, S., Coles, G., Tchuenté, L.A.T., Mbaye, A. & Engels, D. (2009). Praziquantel: its use in control of schistosomiasis in sub-Saharan Africa and current research needs. *Parasitology*, **136**, 1825–1835. 71

Domingos, P. & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, **29**, 103–130. 134

Doyle, M.A., Gasser, R.B., Woodcroft, B.J., Hall, R.S. & Ralph, S.A. (2010). Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics*, **11**, 222+. 23, 60, 146

Duan, Z.H.H., Hughes, B., Reichel, L., Perez, D.M. & Shi, T. (2006). The relationship between protein sequences and their gene ontology functions. *BMC bioinformatics*, **7 Suppl 4**. 22

Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, **7**, e1002195+. 112

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., Westenberger, S.J., Caler, E., Cerqueira, G.C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D.A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J.F.F., de Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A.C.C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R.R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J.L.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O., Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, M.T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P.N., Wickstead, B., Wortman, J., White, O., Fraser, C.M., Stuart, K.D. & Andersson, B. (2005). The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science (New York, N.Y.)*, **309**, 409–415. 68

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E. & Mouse Genome Database Group (2012). The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic acids research*, **40**. 77

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R. & Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, **38**, D211–D222. 86, 88

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, **19**, 99–113. 24

Fleming, A., Chain, E.B. & Florey, H. (1945). Sir Alexander Fleming - Nobel Lecture: Penicillin. 1

Frearson, J.A., Wyatt, P.G., Gilbert, I.H. & Fairlamb, A.H. (2007). Target assessment for antiparasitic drug discovery. *Trends in parasitology*, **23**, 589–595. 6, 10, 61, 103

French, Christopher, T., Lao, Ping, Loraine, Ann, E., Matthews, Brian, T., Yu, Huilan, Dybvig & Kevin (2008). Large-scale transposon mutagenesis of Mycoplasma pulmonis. *Molecular Microbiology*, **69**, 67–76. 32

Friedrich, B.M., Dziuba, N., Li, G., Endsley, M.A., Murray, J.L. & Ferguson, M.R. (2011). Host factors mediating HIV-1 replication. *Virus Research*, **161**, 101–114. 19

Gaiano, N., Amsterdam, A., Kawakami, K., Allende, M., Becker, T. & Hopkins, N. (1996). Insertional mutagenesis and rapid cloning of essential genes in zebrafish. *Nature*, **383**, 829–832. 21

Gallagher, L.A., Ramage, E., Jacobs, M.A., Kaul, R., Brittnacher, M. & Manoil, C. (2007). A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1009–1014. 30, 31, 32

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J.P. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, **40**, D1100–D1107. 5, 9, 73, 81, 90

Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R. & Osterman, A. (2006). Essential genes on metabolic maps. *Current opinion in biotechnology*, **17**, 448–456. 31

Ghosh, A.S., Chowdhury, C. & Nelson, D.E. (2008). Physiological functions of D-alanine carboxypeptidases in Escherichia coli. *Trends in microbiology*, **16**, 309–317. 100

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O. & Venter, J.C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 425–430. 21, 32

Grossmann, S., Bauer, S., Robinson, P.N. & Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, **23**, 3024–3031. 74

Guzman, L.M., Belin, D., Carson, M.J. & Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of Bacteriology*, **177**, 4121–4130. 21

Gwynn, M.N., Portnoy, A., Rittenhouse, S.F. & Payne, D.J. (2010). Challenges of antibacterial discovery revisited: Challenges of antibacterial discovery. *Annals of the New York Academy of Sciences*, **1213**, 5–19. 5

Halgren, T.A. (2009). Identifying and Characterizing Binding Sites and Assessing Druggability. *Journal of Chemical Information and Modeling*, **49**, 377–389. 8, 147

Harborth, J., Elbashir, S.M., Bechert, K., Tuschl, T. & Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. *Journal of cell science*, **114**, 4557–4565. 21

Hasan, S., Daugelat, S., Rao, P.S.S. & Schreiber, M. (2006). Prioritizing Genomic Drug Targets in Pathogens: Application to Mycobacterium tuberculosis. *PLoS Computational Biology*, **2**, e61+. 6

Herrmann, D.J., Peppard, W.J., Ledeboer, N.A., Theesfeld, M.L., Weigelt, J.A. & Buechel, B.J. (2008). Linezolid for the treatment of drug-resistant infections. *Expert review of anti-infective therapy*, **6**, 825–848. 2

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E. & Waterston, R.H. (2005). Genomics in C. elegans: So many genes, such a little worm. *Genome Research*, **15**, 1651–1660. 72

Hillisch, A., Pineda, L.F. & Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discovery Today*, **9**, 659–669. 103

Hitchings, G.H. (1973). Mechanism of Action of Trimethoprim-SulfamethoxazoleI. *Journal of Infectious Diseases*, **128**, S433–S436. 18

Hitchings, G.H. (1989). Nobel lecture in physiology or medicine–1988. Selective inhibitors of dihydrofolate reductase. *In vitro cellular & developmental biology : journal of the Tissue Culture Association*, **25**, 303–310. 18, 103

Høiby, N., Krogh Johansen, H., Moser, C., Song, Z., Ciofu, O. & Kharazmi, A. (2001). Pseudomonas aeruginosa and the in vitroand in vivo biofilm mode of growth. *Microbes and Infection*, **3**, 23–35. 64

Holman, A., Davis, P., Foster, J., Carlow, C. & Kumar, S. (2009). Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi. *BMC Microbiology*, **9**, 243+. 22, 51, 145

Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, **4**, 682–690. 106

Hopkins, A.L. & Bickerton, G.R. (2010). Drug discovery: Know your chemical space. *Nature Chemical Biology*, **6**, 482–483. 5

Hopkins, A.L. & Groom, C.R. (2002). The druggable genome. *Nature reviews. Drug discovery*, **1**, 727–730. 10, 82, 147

Hopkins, A.L., Groom, C.R. & Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today*, **9**, 430–431. 90

Hopkins, A.L., Bickerton, G.R., Carruthers, I.M., Boyer, S.K., Rubin, H. & Overington, J.P. (2011). Rapid analysis of pharmacology for

infectious diseases. *Current topics in medicinal chemistry*, **11**, 1292–1300. 11, 106

Hulsen, T., de Vlieg, J. & Alkema, W. (2008). BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, **9**, 488+. xv, 67

Idsa (2011). IDSA: Facts about Antibiotic Resistance. 2

Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S.M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C.V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R.M., Cronin, A., Cruz, A.K., Davies, R.M., De Gaudenzi, J., Dobson, D.E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A.C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J.C., Müller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'neil, S., Pentony, M., Pohl, T.M., Price, C., Purnelle, B., Quail, M.A., Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J.C., Rutter, S., Saunders, D., Schäfer, M., Schein, J., Schwartz, D.C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D.F., Blackwell, J.M., Stuart, K.D., Barrell, B. & Myler, P.J. (2005). The genome of the kinetoplastid parasite, Leishmania major. *Science*, **309**, 436–442. 68

Jackson, A.P., Sanders, M., Berry, A., McQuillan, J., Aslett, M.A., Quail, M.A., Chukualim, B., Capewell, P., MacLeod, A., Melville, S.E., Gibson, W., Barry, J.D., Berriman, M. & Hertz-Fowler, C.

(2010). The Genome Sequence of Trypanosoma brucei gambiense, Causative Agent of Chronic Human African Trypanosomiasis. *PLoS Negl Trop Dis*, **4**, e658+. 68

Jacob, F. (1977). Evolution and tinkering. *Science*, **196**, 1161–1166. 24

Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., Chun-Rong, L., Guenthner, D., Bovee, D., Olson, M.V. & Manoil, C. (2003). Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14339–14344. xv, 20, 65, 66, 67

Janoir, C., Zeller, V., Kitzis, M.D., Moreau, N.J. & Gutmann, L. (1996). High-level fluoroquinolone resistance in Streptococcus pneumoniae requires mutations in parC and gyrA. *Antimicrobial Agents and Chemotherapy*, **40**, 2760–2764. 10

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, **451**, 990–993+. 2

Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C. & Thornton, J.M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein science : a publication of the Protein Society*, **7**, 233–242. 83

Joubert, F., Harrison, C.M., Koegelenberg, R.J., Odendaal, C.J. & de Beer, T.A. (2009). Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malaria Journal*, **8**, 178+. 6

Kahan, F.M., Kahan, J.S., Cassidy, P.J. & Kropp, H. (1974). The mechanism of action of fosfomycin (phosphonomycin). *Annals of the New York Academy of Sciences*, **235**, 364–386. 100

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P. & Ahringer, J. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231–237. 10, 72, 77

Katara, P., Grover, A., Kuntal, H. & Sharma, V. (2010). In silico prediction of drug targets in Vibrio cholerae. *Protoplasma*, **248**, 799–804. 6

Kawashima, S. & Kanehisa, M. (2000). AAindex: Amino Acid index database. *Nucleic Acids Research*, **28**, 374. 115

Keen, E.C. (2012). Paradigms of pathogenesis: targeting the mobile genetic elements of disease. *Frontiers in Cellular and Infection Microbiology*, **2**. 19

Keller, T.H., Pichota, A. & Yin, Z. (2006). A practical view of 'druggability'. *Current Opinion in Chemical Biology*, **10**, 357–361. 7

Kempheus, K. (2005). Essential genes. *WormBook*. 21

Kerr, K.G. & Snelling, A.M. (2009). Pseudomonas aeruginosa: a formidable and ever-present adversary. *The Journal of hospital infection*, **73**, 338–344. 64

Khattab, M.A. (2009). Targeting host factors: A novel rationale for the management of hepatitis C virus. *World Journal of Gastroenterology*, **15**, 3472–3479. 19

Klein, B., Tenorio, E., Lazinski, D., Camilli, A., Duncan, M. & Hu, L. (2012). Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. *BMC Genomics*, **13**, 578+. 61, 145

Kovalevskaya, N.V., Smurnyy, Y.D., Polshakov, V.I., Birdsall, B., Bradbury, A.F., Frenkiel, T. & Feeney, J. (2005). Solution structure of human dihydrofolate reductase in its complex with trimethoprim and NADPH. *Journal of biomolecular NMR*, **33**, 69–72. 102, 103

Kovalevskaya, N.V., Smurnyi, E.D., Birdsall, B., Feeney, J. & Polshakov, V.I. (2007). Structural factors determining the binding selectivity of

the antibacterial drug trimethoprim to dihydrofolate reductase. **41**, 350–353. 103

Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S. & Brenk, R. (2011). DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *Journal of chemical information and modeling*, **51**, 2829–2842. 8

Kruger, F., Rostom, R. & Overington, J. (2012). Mapping small molecule binding data to structural domains. *BMC Bioinformatics*, **13**, S11+. 147

Kumar, S., Chaudhary, K., Foster, J.M., Novelli, J.F., Zhang, Y., Wang, S., Spiro, D., Ghedin, E. & Carlow, C.K.S. (2007). Mining Predicted Essential Genes of Brugia malayi for Nematode Drug Targets. *PLoS ONE*, **2**, e1189+. 6, 23

Kuntz, I.D., Chen, K., Sharp, K.A. & Kollman, P.A. (1999). The maximal affinity of ligands. *Proceedings of the National Academy of Sciences*, **96**, 9997–10002. 90

L. Hopkins, A., Richard Bickerton, G., M. Carruthers, I., K. Boyer, S., Rubin, H. & P. Overington, J. (2011). Rapid Analysis of Pharmacology for Infectious Diseases. *Current Topics in Medicinal Chemistry*, **11**, 1292–1300. 5, 7

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt,

A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler,

G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A. & Morgan, M.J. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. 110

Langridge, G.C., Phan, M., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., Wain, J., Parkhill, J. & Turner, A.K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, **19**, 2308–2316. 21

Le Guilloux, V., Schmidtke, P. & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, **10**, 1–11. 8

Li, L., Stoeckert, C.J. & Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**, 2178–2189. 33, 35

Liberati, N.T., Urbach, J.M., Miyata, S., Lee, D.G., Drenkard, E., Wu, G., Villanueva, J., Wei, T. & Ausubel, F.M. (2006). An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 2833–2838. xv, 65, 66, 67

Livermore, D.M., Blaser, M., Carrs, O., Cassell, G., Fishman, N., Guidos, R., Levy, S., Powers, J., Norrby, R., Tillotson, G., Davies, R., Projan, S., Dawson, M., Monnet, D., Keogh-Brown, M., Hand, K., Garner, S., Findlay, D., Morel, C., Wise, R., Bax, R., Burke, F., Chopra, I., Czaplewski, L., Finch, R., Livermore, D., Piddock, L.J.V., White, T. & on behalf of the British Society for Antimicrobial Chemotherapy Working Party on The

Urgent Need: Regenerating Antibacterial Drug Discovery and Development (2011). Discovery research: the scientific challenge of finding new antibiotics. *Journal of Antimicrobial Chemotherapy*, **66**, 1941–1944. 4, 5

Lopez, P., Espinosa, M., Greenberg, B. & Lacks, S.A. (1987). Sulfonamide resistance in Streptococcus pneumoniae: DNA sequence of the gene encoding dihydropteroate synthase and characterization of the enzyme. *Journal of Bacteriology*, **169**, 4320–4326. 18

Magrane, M. & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009. 112

Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002). The Protein Kinase Complement of the Human Genome. *Science*, **298**, 1912–1934. 110

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D. & Bryant, S.H. (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic acids research*, **35**, D237–D240. 86

Martin, A.C., Toda, K., Stirk, H.J. & Thornton, J.M. (1995). Long loops in proteins. *Protein engineering*, **8**, 1093–1101. 89

Martin, D.M.A., Miranda-Saavedra, D. & Barton, G.J. (2009). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Research*, **37**, D244–D250. 112

Mayor, L.R., Fleming, K.P., Müller, A., Balding, D.J. & Sternberg, M.J. (2004). Clustering of Protein Domains in the Human Genome. *Journal of Molecular Biology*, **340**, 991–1004. 106

Metz, J.T., Johnson, E.F., Soni, N.B., Merta, P.J., Kifle, L. & Hajduk, P.J. (2011). Navigating the kinome. *Nature Chemical Biology*, **7**, 200–202. 128

Micheletti, C. & Orland, H. (2009). MISTRAL: a tool for energy-based multiple structural alignment of proteins. *Bioinformatics (Oxford, England)*, **25**, 2663–2669. 108

Miranda-Saavedra, D. & Barton, G.J. (2007). Classification and functional annotation of eukaryotic protein kinases. *Proteins*, **68**, 893–914. 75

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. (1998a). JOY: protein sequence-structure representation and analysis. *Bioinformatics.*, **14**, 617–623+. 108

Mizuguchi, K., Deane, C.M., Blundell, T.L. & Overington, J.P. (1998b). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein science : a publication of the Protein Society*, **7**, 2469–2471. 108

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, **247**, 536–540. 85

Mushegian, A.R. & Koonin, E.V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10268–10273. 23, 24, 50

Nagaraj, N.S. & Singh, O.V. (2010). Using Genomics to Develop Novel Antibacterial Therapeutics. *Critical Reviews in Microbiology*, **36**, 340–348. 4

Orengo, C.A. (1999). CORA–topological fingerprints for protein structural families. *Protein science : a publication of the Protein Society*, **8**, 699–715. 108

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, **5**, 1093–1108. 85

Ortí, L., Carbajo, R.J., Pieper, U., Eswar, N., Maurer, S.M., Rai, A.K., Taylor, G., Todd, M.H., Pineda-Lucena, A., Sali, A. & Marti-Renom, M.A. (2009). A Kernel for Open Source Drug Discovery in Tropical Diseases. *PLoS Neglected Tropical Diseases*, **3**, e418+. 104

O'Shea, R. & Moser, H.E. (2008). Physicochemical Properties of Antibacterial Compounds: Implications for Drug Discovery. *Journal of Medicinal Chemistry*, **51**, 2871–2878. 5

Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, **5**, 993–996+. 82, 93

Paolini, G.V., Shapland, R.H.B., van Hoorn, W.P., Mason, J.S. & Hopkins, A.L. (2006). Global mapping of pharmacological space. *Nature Biotechnology*, **24**, 805–815. 5

Payne, D.J., Gwynn, M.N., Holmes, D.J. & Rosenberg, M. (2004). *Genomic Approaches to Antibacterial Discovery*, vol. 266, chap. 11, 231–259. Humana Press, Totowa, NJ. 4

Payne, D.J., Gwynn, M.N., Holmes, D.J. & Pompliano, D.L. (2006). Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Reviews Drug Discovery*, **6**, 29–40. 4, 147

Peacock, C.S., Seeger, K., Harris, D., Murphy, L., Ruiz, J.C., Quail, M.A., Peters, N., Adlem, E., Tivey, A., Aslett, M., Kerhornou, A., Ivens, A., Fraser, A., Rajandream, M.A.A., Carver, T., Norbertczak, H., Chillingworth, T., Hance, Z., Jagels, K., Moule, S., Ormond, D., Rutter, S., Squares, R., Whitehead, S., Rabbinowitsch, E., Arrowsmith, C., White, B., Thurston, S., Bringaud, F., Baldauf, S.L., Faulconbridge, A., Jeffares, D., Depledge, D.P., Oyola, S.O., Hilley, J.D., Brito, L.O., Tosi, L.R., Barrell, B.,

Cruz, A.K., Mottram, J.C., Smith, D.F. & Berriman, M. (2007). Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature genetics*, **39**, 839–847. 68

Pepin, J., Guern, C., Milord, F. & Schechter, P.J. (1987). Difluoromethylornithine for arseno-resistant trypanosoma brucei gambiense sleeping sickness. *The Lancet*, **330**, 1431–1433. 105

Phan, I.Q., Pilbout, S.F., Fleischmann, W. & Bairoch, A. (2003). NEWT, a new taxonomy portal. *Nucleic acids research*, **31**, 3822–3823. 27

Pieper, U., Eswar, N., Webb, B.M., Eramian, D., Kelly, L., Barkan, D.T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M.A., Davis, F.P. & Sali, A. (2009). MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, **37**, D347–D354. 104

Projan, S.J. & Bradford, P.A. (2007). Late stage antibacterial drugs in the clinical pipeline. *Current opinion in microbiology*, **10**, 441–446. 2

Prudencio, M. & M. Mota, M. (2012). Targeting Host Factors to Circumvent Anti-Malarial Drug Resistance. *Current Pharmaceutical Design*, **19**, 290–299. 19

Pullan, R. & Brooker, S. (2008). The health impact of polyparasitism in humans: are we under-estimating the burden of parasitic diseases? *Parasitology*, **135**, 783–794+. 12

Raman, K., Yeturu, K. & Chandra, N. (2008). targetTB: A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology*, **2**, 109+. 6

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, **16**, 276–277. 88

Rizzi, C., Frazzon, J., Ely, F., Weber, P.G., da Fonseca, I.O., Gallas, M., Oliveira, J.S., Mendes, M.A., de Souza, B.M., Palma, M.S., Santos, D.S. & Basso, L.A. (2005). DAHP synthase from Mycobacterium tuberculosis H37Rv: cloning, expression, and purification of functional enzyme. *Protein Expression and Purification*, **40**, 23–30. 100

Robinson, K.A. & Beverley, S.M. (2003). Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. *Molecular and Biochemical Parasitology*, **128**, 217–228. 21

Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W.J., Davis, P., Fernandes, J., Fiedler, T.J., Han, M., Harris, T.W., Kishore, R., Lee, R., McKay, S., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Schindelman, G., Schwarz, E.M., Spooner, W., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J. & Sternberg, P.W. (2008). WormBase 2007. *Nucl. Acids Res.*, **36**, D612–617. 72

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., Chen, W.J., Pu, F., Peng, Y., Li, J., Xi, F., Li, S., Li, Y., Zhang, Z., Yang, X., Zhao, M., Wang, P., Guan, Y., Cen, Z., Zhao, X., Christner, M., Kobbe, R., Loos, S., Oh, J., Yang, L., Danchin, A., Gao, G.F., Song, Y., Li, Y., Yang, H., Wang, J., Xu, J., Pallen, M.J., Wang, J., Aepfelbacher, M., Yang, R. & Consortium, E.c.O.G.A.C.S. (2011). Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *The New England Journal of Medicine*, **365**, 718–724+. 4

Rohmer, L., Hocquet, D. & Miller, S.I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends in microbiology*, **19**, 341–348. 31

Russell, R.B. & Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins*, **14**, 309–323. 108

Saldanha, A.J. (2004). Java Treeviewextensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248. xxii, 124, 181

Schreyer, A. & Blundell, T. (2009). CREDO: A ProteinLigand Interaction Database for Drug Discovery. *Chemical Biology & Drug Design*, **73**, 157–167. 110, 111, 140

Schwegmann, A. & Brombacher, F. (2008). Host-Directed Drug Targeting of Factors Hijacked by Pathogens. *Science Signaling*, **1**, re8. 19

Sheinerman, F.B., Giraud, E. & Laoui, A. (2005). High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *Journal of molecular biology*, **352**, 1134–1156. 104, 116

Shiau, A.K., Barstad, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A. & Greene, G.L. (1998). The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell*, **95**, 927–937. 93

Siddiqui, A.S., Dengler, U. & Barton, G.J. (2001). 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201. 83

Silver, L.L. (2011). Challenges of antibacterial discovery. *Clinical microbiology reviews*, **24**, 71–109. 13, 20, 52

Singh, N.K., Selvam, S.M. & Chakravarthy, P. (2006). T-iDT : tool for identification of drug target in bacteria and validation by Mycobacterium tuberculosis. *In Silico Biology*, **6**, 485–493+. 6

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**, 195–197. 88

Spellberg, B. (2008). Dr. William H. Stewart: Mistaken or Maligned? *Clinical Infectious Diseases*, **47**, 294. 1

Thompson, J. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882. xx, 135

Todar, K. (2008). *Bacterial Resistance to Antibiotics*. 1

van Oosterom, A.T., Judson, I., Verweij, J., Stroobants, S., di Paola, E.D., Dimitrijevic, S., Martens, M., Webb, A., Sciot, R. & Van Glabbeke, M. (2001). Safety and efficacy of imatinib (STI571) in metastatic gastrointestinal stromal tumours: a phase I study. *The Lancet*, **358**, 1421–1423. 105

van Westen, G.J.P., Wegner, J.K., Geluykens, P., Kwanten, L., Vereycken, I., Peeters, A., IJzerman, A.P., van Vlijmen, H.W.T. & Bender, A. (2011). Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. *PLoS ONE*, **6**, e27518+. 116, 117

Vaudaux, P., Pittet, D., Haeberli, A., Huggler, E., Nydegger, U.E., Lew, D.P. & Waldvogel, F.A. (1989). Host Factors Selectively Increase Staphylococcal Adherence on Inserted Catheters: A Role for Fibronectin and Fibrinogen or Fibrin. *Journal of Infectious Diseases*, **160**, 865–875. 19

Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. & Henrick, K. (2005). E-MSD: an integrated data resource for bioinformatics. *Nucleic acids research*, **33**, D262–265. 110

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R.,

Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman,

B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001). The Sequence of the Human Genome. *Science*, **291**, 1304–1351. 110

Volkamer, A., Kuhn, D., Rippmann, F. & Rarey, M. (2012). DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics (Oxford, England)*, **28**, 2074–2075. 8

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. & Barton, G.J. (2009). Jalview Version 2a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191. xvii, 109

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, **28**, 31–36. 129

Weiss, B., Davidkova, G. & Zhou, L.W. (1999). Antisense RNA gene therapy for studying and modulating biological processes. *Cellular and molecular life sciences : CMLS*, **55**, 334–358. 21

White, T.A. & Kell, D.B. (2004). Comparative Genomic Assessment of Novel Broad-Spectrum Targets for Antibacterial Drugs. *Comparative and Functional Genomics*, **5**, 304–327. 4, 6

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., Bakkoury, M.E., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N.,

Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Véronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M. & Davis, R.W. (1999). Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science*, **285**, 901–906. 10

Wolf, John, E., Shander, Douglas, Huber, Ferdinand, Jackson, Joseph, Lin, Chen-Sheng, Mathes, Barbara, M., Schrode & Kathy (2007). Randomized, double-blind clinical evaluation of the efficacy and safety of topical eflornithine HCl 13.9 cream in the treatment of women with facial hair. *International Journal of Dermatology*, **46**, 94–98. 105

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N. & Suzek, B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, **34**, D187–D191. 69

Wyatt, P.G., Gilbert, I.H., Read, K.D. & Fairlamb, A.H. (2011). Target validation: linking target and chemical properties to desired product profile. *Current topics in medicinal chemistry*, **11**, 1275–1283. 6, 68

Xu, P., Ge, X., Chen, L., Wang, X., Dou, Y., Xu, J.Z., Patel, J.R., Stone, V., Trinh, M., Evans, K., Kitten, T., Bonchev, D. & Buck, G.A. (2011). Genome-wide essential gene identification in Streptococcus sanguinis. *Scientific reports*, **1**. 61, 145

Yonath, A. (2005). ANTIBIOTICS TARGETING RIBOSOMES: Resistance, Selectivity, Synergism, and Cellular Regulation. *Annual Review of Biochemistry*, **74**, 649–679. 52

Yuan, Y., Xu, Y., Xu, J., Ball, R.L. & Liang, H. (2012). Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics*, **28**, 1246–1252. 146

Zhang, R., Ou, H.Y.Y. & Zhang, C.T.T. (2004). DEG: a database of essential genes. *Nucleic acids research*, **32**, D271–D272. 27

Zoeiby, A.E., Sanschagrin, F., Darveau, A., Brisson, J.R. & Levesque, R.C. (2003). Identification of novel inhibitors of Pseudomonas aeruginosa MurC enzyme derived from phage-displayed peptide libraries. *Journal of Antimicrobial Chemotherapy*, **51**, 531–543. 100

Zuccotto, F., Zvelebil, M., Brun, R., Chowdhury, S.F., Di Lucrezia, R., Leal, I., Maes, L., Ruiz-Perez, L.M., Gonzalez Pacanowska, D. & Gilbert, I.H. (2001). Novel inhibitors of Trypanosoma cruzi dihydrofolate reductase. *European Journal of Medicinal Chemistry*, **36**, 395–405. 103